A TABLET + AUGMENTED REALITY INTERFACE FOR INTERACTIVE MULTIPLE LINEAR REGRESSION WITH GEOSPATIAL DATA

by

Sathaporn Hu

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
January 2024

This thesis is dedicated to those who have faith in me, even if I might not be worthy of it. This work uses software invented by Ukrainians. Sláva Ukrayíni!

Table of Contents

List of Tables

List of Figures

Abstract

AR headsets support mobile large-area displays which make them suitable for immersive analytics. We explore using Microsoft HoloLens v2, an augmented reality headset, for immersive geospatial analyses. Since HoloLens are limited in terms of display quality and input, we explored enhancing the devices with computer tablets. We created Gander, a geospatial AR+tablet application capable of multiple linear regression. Gander uses Stacked glyphs for visualization; Stacked is a glyph composition technique which uses the parallax technique to allow the user to compose and decompose coloured glyphs just by changing their viewing angle. Where the glyphs overlap, the colour could be blended to express other information. We conducted three studies. Each study focused on different aspects of Gander. The first study was about understanding how users navigate through large room-sized fields of glyphs as rendered by Gander. We compared two types of glyph-based visualizations: Polyline and Radial. Polyline is a shape-based multivariate glyph visualization technique. Radial is a technique similar to Stacked, but the parallax effect has been disabled and the glyphs are arranged radially. We did not use Stacked in the study to avoid introducing 3D as a confounding variable. The results show that Polyline induced more tablet-based panning movements while Radial encouraged more gaze movement. The second study compared Stacked against Radial. It supplemented the first study by involving the parallax effect. For each trial, a participant observed individual glyph composite and indicated their values. We found Stacked was helpful in terms of speed when the glyphs were far away, because the glyphs were already decomposed. In terms of accuracy, there were other external factors (e.g., colourmap, number of glyphs in a composite, visual aid) more important than the parallax effect. The third study was a walkthrough demonstration study. Experts were interviewed and interacted with Gander. Since we engaged in design as research, we designed Gander through the first principle analysis of the literature. This study showed that Gander required customizability and more interactivity. In the end, we created a low-fidelity prototype for the updated version of Gander, using the results of the three studies.

List of Abbreviation and Symbols Used

α	Threshold of Type I Error probability.
AbsDiff	Absolute Difference
ANOVA	Analysis of Variance
AI	Artificial Intelligence
AIC	Akaike Information Criterion
AR	Augmented Reality
AR-HWD	Augmented Reality Head-worn Display
ART-ANOVA	Aligned Rank Transformation Analysis of Variance
ART-Contrasts	The contrast tests for Aligned Rank Transformation Analysis of Variance
bSOUS	bold Sign Of the UnSeen
Ch	Chapter
CVD	Colour-vision Deficiency
cm	Centimetre
χ^2	Chi-squared distribution/test
d	Cohen's d (Effect size)
df	Degree of Freedom
DV	Dependent Variable
EDA	Exploratory Data Analysis
E_L	Effect size for likelihood ratios standardized to be between 0 and 1.

ESQ	Effect Size Questionnaire
η^2	Eta-squared (Effect size)
F	F-statistic
Fig	Figure
F+C	Focus+Context
FOV	Field-of-view
fSOUS	faint Sign Of the UnSeen
GCS	Glyph Composition Size
GIS	Geographic Information System
GLM	Generalized Linear Model
GPS	Global Positioning System
GWR	Geographically Weighted Regression
Γ	Gamma distribution
HCI	Human-computer Interaction
HWD	Head-worn Display
in	Inch
IML	Interactive Machine Learning
IQ	Intelligence Quotient
IV	Independent Variable
JSON	JavaScript Object Notation
Kendall's W	Kendall's Coefficient of Concordance
Kerby's r	Kerby's Rank biserial correlation (Not Pearson's correlation.)

LISA	Local Indicators of Spatial Autocorrelation
log	Logarithm
m	Metre
m^2	Squared Metre
mm	Millimetre
MAR	Mobile Augmented Reality
MD	Sample median
	MiRA Mixed Reality Agent
ML	Machine Learning
MLR	Multiple Linear Regression
MR	Mixed Reality
ms	Milliseconds
MSD	Mean Squared Displacement
NASA-TLX	NASA Task Load Index
Obj	Research Objective
OLS	Ordinary Least Square
OST	Optical See-through
OST-HWD	Optical See-through Head-worn Display
O+D	Overview+Detail
p	Probability of Type I Error
px	Pixels
PCA	Principle Component Analysis

PERMANOVA	Permutational Multivariate Analysis of Variance
PoI	Point of Intersection
ψ_1	Trigamma Function
R^2	R-squared (Effect size)
R^2_{GLMM}	Nakagawa's R-squared for mixed-effect models (Effect size)
RGB	Red-Blue-Green colour system
sd	Standard Deviation
SE	Standard Error
Sec	Section
SEM	Spatial Error Model
SLM	Spatial Lag Model
SNARC	Spatial-numerical Association of Response Codes
SUS	System Usability Scale
t	t-test Statistic
UMUX	Usability Metric for User Experience
USA	United States of America
VLAT	Visualization Literacy Assessment Test
VIF	Variance Inflation Factors
VR	Virtual Reality
VST	Video See-through
VST-HWD	Video See-through Head-worn Display
$\bar{x}$	Mean
XAI	Explainable Artificial Intelligence

Acknowledgements

My work received financial support from multiple Canadian governmental organizations namely, Mitacs, Dalhousie University, Natural Sciences and Engineering Research Council of Canada (NSERC), and the Government of Nova Scotia. Global Artificial Intelligence Accelerator (GAIA), a research unit at Ericsson, Montreal partially funded my work with Mitacs. Through my cooperation with GAIA, I was able to file a patent based on the prototype described in this thesis.

I would like to thank Prof. Derek Reilly for his patience, kindness, and advice throughout the program. Furthermore, I would like to extend my thanks to the members of the examination committee: Prof. Jamie Blustein, Prof. Joseph Malloch, and Prof. Fernando Paulovich. I must also extend my gratitude toward Prof. Pourang Irani, and Prof. Rita Orji who were my external examiners. Other professors who provided support and encouragement were: Prof. Miguel Garcia-Ruiz, Prof. Kirstie Hawkey, Prof. Bonnie Mackay, Dr. Jill McSweeney-Flaherty, Prof. Aaron Newman, Prof. Rina Webhe, and Prof. Thomas Trappenburg.

I acknowledge the support from the members of GEMLab. Some members of the laboratory assisted me in creating manuscripts whose materials were incorporated into this thesis. The members were: Hariprasanth Devisagammani, Muhammed Raza, and Ramanpreet Kaur. Furthermore, some of my work was supervised by Prof. Reilly and Prof. Malloch. Since this thesis was written using materials partially developed by the noted individuals, I have decided to use “we” instead of “I” in the abstract and for most of the thesis.

In addition to the professors and lab members, I would like to thank other individuals. First, I would like to thank my girlfriend Helen Dow and my family for the support that I received throughout my program. Secondly, I would like to acknowledge my friendship with Chinenye Ndulue, Zachary O’Keefe, and Justin Hartherly. They helped me to settle in Halifax at the beginning of the program. Thirdly, I must extend my gratitude towards Dr. Saman Bashbaghi, a researcher at Ericsson who essentially acted as my second supervisor during my Mitacs internship. Fourthly, I would like to thank Jasmin Kaur, Sai Pavan, and Tejas Patel who were my Hackathon teammates. While I designed all aspects of AirSeer (a precursor to Gander) at the Hackathon,

they helped to present the prototype. Fifthly, I would like to thank Calyssa Skeggs for her assistance during the later stage of my program. Lastly and most importantly, I cannot forget to thank the anonymous individuals who participated in my studies.

Chapter 1

Introduction

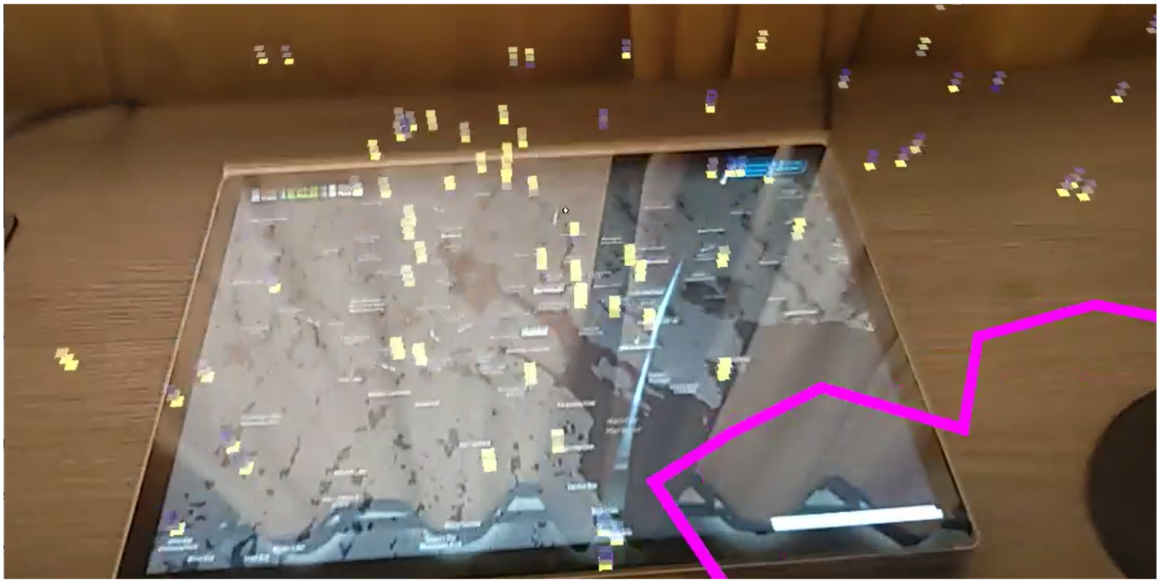


Figure 1.1: A prototype of Gander. Screenshot recorded on Microsoft HoloLens v2.

Figure 1.2: Microsoft HoloLens v2–the augmented reality device that we use in our work. The image is from Wikimedia [Kcida10, 2015].

We designed and implemented Gander, a geospatial analysis system with an augmented reality (AR) + tablet interface. We named the system after Gander, Canada and the expression: "To take a gander." Gander uses lightweight 3D visualization to present geospatial data, and likelihoods of models. AR glyphs are mounted on top and around the tablet device. The tablet itself renders a map in the background. Through tablet-based swipe gestures, the user pans the AR content through swipe gestures. In addition to panning, the tablet contains a menu that allows the user to manipulate the AR-based visualization, and to create a statistical model.

Gander takes advantage of mobility afforded by an untethered optical see-through head-worn display (OST-HWD). Such a display device supports presenting large virtual content while allowing the user to still freely move around [Pavanatto et al., 2021]. FieldView by Whitlock et al. [2020] is an example of an AR-based geospatial analysis application that takes advantage of increased mobility. Their software, a mobile phone + AR system, allows the user to collect and analyze data in the field, and to create large immersive visualizations that incorporate spatial information in the physical world. AR, alongside other mixed reality (MR) technologies, plays a key role in implementing immersive analytics (IA), which Dwyer et al. [2018] define as: “the use of engaging, embodied analysis tools to support data understanding and decision making.” One possible use of IA is the implementation of Digital Earth. Çöltekin et al. [2020] define Digital Earth as an ultimate and idealistic goal of geospatial software that essentially gives the user a virtual copy of Earth to explore in any manner that they like.

While AR is helpful, it has multiple limitations. First, as pointed out by Feiner and Shamash [1991], an AR input system or device like an AR controller wand can be less accurate than another input device like mice. Using an AR input system in conjunction with a separate piece of hardware can improve accuracy [Feiner and Shamash, 1991]. Although some OST-HWDs have additional input methods (e.g. gaze [Kress and Cummings, 2017] and hand gestures [Kress and Cummings, 2017, Pulli, 2017]), these are still inaccurate and require additional physical input devices [Soares et al., 2021]. In addition to the input limitation, OST-HWDs have poor display resolutions [Itoh et al., 2021] which may require additional display monitors. Using a tablet solves the issues of input inaccuracy and low display resolution at once. A tablet with good multi-touch support allows the user to use touch gestures to make accurate input. Meanwhile, the tablet’s screen can display details that the AR system cannot. Our research involves studying how a tablet, when used as an input and a display device, impacts the overall usability of an AR+tablet interface.

There are many areas of possible applications for an AR+tablet interface. For instance, MARVIS by Langner et al. [2021] uses them for 3D collaborative visual analytics. Our work focuses on geospatial analysis for multiple reasons. First, geospatial analysis has a wide range of relevant real-world applications that are often high-stakes. For instance, Snow, in the canonical example that combines data visualization and health, used geospatial analysis and visualizations to shut down a contaminated water source during a cholera outbreak in 1854 [Brody et al., 2000]. Another example is from one of our studies where participants indicated that their works were used to influence maritime policies. Secondly, a geospatial analysis benefits from the use of various types of visualization to either make an inference and to communicate the results. For instance, Kumar et al. [2013] render predicted values of a regression model as a raster overlaying a map of a rain forest. The raster allows them to infer variations of the biomass in the forest, to communicate the predicted values for various people. With AR, the visualization could be expanded and manipulated in various ways. Tangible Globes by Satriadi et al. [2022] allow the user to use a spherical control to alter AR-based visualizations for many possible use cases. They can use the system to easily create large-area, room-sized visualizations. Large AR visualizations are good for geospatial analysis results for many people at once. An example of this is Airbus Tactical Sandbox which allows multiple war planner to plan a combat mission using AR representations of the battlefield [Walsh et al., 2023]. Thirdly, geospatial analysis dovetails into Digital Earth, or a creation of a virtual copy of Earth that enables the new mode of interactions and inferences [Çöltekin et al., 2020]. Çöltekin et al. [2020] argue Digital Earth requires advances in immersive mixed reality technologies. While our work is not about Digital Earth, it helps to make progress toward the goal.

Geospatial information can be visualized in many ways. We chose glyph-based visualization for Gander, because we want the user to consider information at the level of individual points. Most statistical analyses generate aggregated information (e.g. mean, coefficients). While they are useful for generalization, they can hide patterns from users. For instance, a cartogram can hide local information of the whole area, making the user unaware of the local trends. This can lead to an ecological fallacy where aggregated information (e.g., mean, median) is used to make a general statement without any regard to individual data point [Salkeld and Antolin, 2020].

We conducted three studies to evaluate different aspects of Gander. We name the three studies: the synoptic study, the elementary study, and the walkthrough demonstration study. The synoptic study and the elementary studies focused on the perceptual aspects of visualization with an AR+tablet interface. Each study used

Techniques	**Synoptic Study**	**Elementary Study**	**Walkthrough Demonstration Study**
Polyline	X		
Stacked		X	X
Radial	X	X	

Table 1.1: A table summarizing how Polyline, Stacked, and Radial were used in the studies. X denotes that the technique was used.

different sets of glyph-based visualization techniques. Please refer Tab. 1.1 for the names of the technique used in the studies.

The synoptic study compares glyph fields created using two visualization techniques: Polyline [Opach et al., 2018], and Radial. Polyline is a small multiples technique based on line charts. Meanwhile, Radial are coloured square glyphs composited in a radial pattern. To increase the external validity of the synoptic study, we used semi-naturalistic study tasks with real-world data. We categorized our tasks as: pre-fit, and post-fit. The pre-fit tasks involved assessing data for model creation while the post-fit tasks were about assessing the qualities of multiple models. The synoptic study is not a 3D visualization study. Instead, it focuses on how shape and colour visual channels affect map panning, gaze trajectory, and navigation around glyph fields. We found Polyline tended to induce more tablet panning while Radial tended to elicit more gaze-based scanning.

The elementary study is a visualization study that explores the light use of 3D visualization through the parallax effect. We compared Radial and Stacked compositions to see if a 3D arrangement has any effect. Stacked composes glyphs by making them float on top of each other—allowing the user to compose and decompose the glyphs through the parallax effect (see [Rouan, 2015]). We found that the parallax effect can speed up glyph comprehension when the glyphs are further away–because the glyphs already appear decomposed to the user. Furthermore, this study reveals the importance of glyph syntax—or how glyphs were arranged together. As we add more glyphs into a composite, the composite becomes harder to comprehend which slows down the user and makes them less accurate. The study also provides some insights into colourmap designs.

The walkthrough demonstration study involved interviews with experts with spatial analysis backgrounds and a demonstration of the prototype. We learned that experts tend to use spatial analysis tools very differently from each other. Therefore, a user may wish to customize Gander to suit their own usage. Furthermore, Gander should provide overview information (e.g., statistical tables) alongside the glyph-based visualization. The participants also suggest reducing steps for certain tasks (e.g., switching glyph layers on the screen).

Overall, we found that there visual channels can affect how the user perceives and interact with AR+tablet interfaces. However, superior instruments are necessary to truly understand how the user makes inferences with Gander. Furthermore, the results of the walkthrough demonstration study provides opportunities for design changes.

1.1 Research Objectives

Our research objectives revolve around designing an IML system for AR+Tablet, and evaluating it with human-computer interaction (HCI) methodologies. Since Gander primarily uses glyph-based visualization, our studies emphasize glyph comprehension and how the user may interact with a glyph field.

1.1.1 Obj 1: Design as research with a vertical slice prototype of geospatial analysis tool for an AR+tablet hybrid user interface

Combining a mobile device and AR has been shown to be beneficial for geospatial analysis [Whitlock et al., 2020, Satriadi et al., 2022]. However, prior literature is more focused on data collection [Whitlock et al., 2020], and visualization [Whitlock et al., 2020, Satriadi et al., 2022]. As Dudley and Kristensson [2018] point out, statistical analysis tasks do not simply end with visualization; some researchers still need to create statistical models, and assess their quality. We create two terms: “pre-fit” and “post-fit.” The former refers to actions performed before data fitting is complete while the latter refers to actions for assessing the created model.

Knowing what a geospatial analyst does during the pre-fit and post-fit stages is difficult, because there are many possible statistical practices. To understand this, we can first look at two types of spatial analysis techniques: geographically weighted regression (GWR), and spatial error models. The first one involves multiple rounds of regressions and the resulting model is a hierarchical one [Comber et al., 2023]. Meanwhile, a spatial error model simply uses the spatial information as an error term [Zhang et al., 2009]. These two techniques demand very different types of analysis.

To understand how a typical geospatial analyst operates, we engage in "design-as-research" which is a process where we develop a prototype to better understand the research space [Stapleton, 2005]. Stapleton [2005] used this process when he was trying to research into using games as a way to teach physics. He found that game design was a nebulous concept, and to understand and research it, a game must first be designed. Like Stapleton's case, despite our efforts researching in multiple areas (e.g., early development of statistics [Ziliak, 2008, 2019], interactive machine learning [Dudley and Kristensson, 2018], the New Statistics [Cumming, 2014]), we realize that to truly understand how statistical work is performed is to design software that supports it.

Our design process uses vertical slicing. Vertical slicing, according to Ratner and Harvey [2011], is a process where we create a functioning prototyping that allows a user to start and finish the main task. The prototype has limited functionalities. Gander is a vertical slice because it allows the user to start from raw data and end with a multiple linear regression (MLR) model. However, it does not allow the user to filter the data or choose other types of geospatial analysis techniques. A vertical slice is beneficial, because it showcases how the final prototype may behave. An alternative approach is to design and test software in a piecemeal fashion. Ratner and Harvey [2011] argue that it can be difficult for observers to understand how all pieces will be combined into a single piece of software.

1.1.2 Obj 2: Designing and evaluation of an appropriate glyph-based visualization

Glyph-based visualization can provide granular information to the user. This is in contrast with other map-based visualizations like the cartogram which presents overview or aggregated information of an area. We are not eschewing representation of aggregated information. Rather, we think that in order to create an effective visualization of aggregated information, we must first understand how the user gleans information from individual data points. Ropinski et al. [2011], Borgo et al. [2013], Fuchs et al. [2017] state that there are many ways to customize the appearance of glyphs. Borgo et al. [2013] calls an appearance attribute a visual channel, and each attribute has separate different sets of benefits and disadvantages. Our work mainly focuses on the colour channel, because it is good for providing information at the pre-attentive phase [Ropinski et al., 2011].

While colours are worse for closer examination [Ropinski et al., 2011] and OST-HWDs distort them [Itoh et al., 2021], they do not encounter limitations that other visual channels have. Size can be affected by overdrawing (i.e. glyphs overlapping so much that the user can have difficulty understanding them [Mayorga and Gleicher, 2013]); therefore, larger glyphs can become less comprehensible in a dense glyph field. Shape can impact devices with limited graphical capabilities like Microsoft HoloLens v2. Earlier, our glyphs were circular. However, the circular shapes had too vertices for a Microsoft HoloLens v2 to render properly. Like shapes, textures may also be difficult to render.

We developed two techniques: Radial and Stacked. Both techniques rely on small coloured squares. Since each square can only express univariate data, the techniques compose multiple squares into multivariate composites. Radial refers to the squares being arranged radially. Since the squares lie on the same plane, this technique has 2D dimensionality. Meanwhile, Stacked glyphs are floating on top of each other, making the technique 3D. The user can change their viewing angle to compose and decompose a Stacked composite through the parallax effect. For both techniques, where the glyphs appear to be overlapping, other information can be expressed. For instance, we can multiply the colours of two glyphs to convey multiplicative information.

To assess the effect of colours on glyph fields, we compared our colour-based technique against a control technique, Polyline [Opach et al., 2018] in the synoptic study. Since the square colour glyphs were composited in a radial pattern, we called the technique Radial. Polyline is a multivariate small multiple technique (i.e. miniature version of an existing visualization technique [van den Elzen and van Wijk, 2013]) based on line charts. One can argue that it is a shape-based technique, because it expresses its values through zig-zagging lines. Since we can implement Polyline using line renderers and not 3D objects, a HoloLens v2 can render Polyline glyphs without any performance issues. This would not be true for other types of small multiples, like the ones based on bar charts. We found that Polyline induced more tablet scrolling as the participants wanted to bring them closer. Meanwhile, the user tended to view Radial glyphs from afar.

Since Polyline is a 2D technique, we did not directly compare it against Stacked in the synoptic study. Had we made the comparison, Stacked's ability to compose or decompose through the parallax effect would have served as a significant confounding variable. Instead, Stacked and Radial were only compared in the elementary study which we specifically designed to test Stacked's ability. Unlike the synoptic study, the participants only judged one composite at a time. This allowed us to better measure the effectiveness of the techniques in terms of accuracy and time. We found that the distinction between 2D and 3D did not play much difference in terms of effectiveness (i.e. accuracy and time). Instead, the arrangement of the glyphs and the numbers of glyphs in composites played more role. Although this study did not compare multiple colourmaps, the results indicate additional researches in that direction. In this study, we followed the best practice which involves using a divergent colourmap for value-judgment tasks (i.e. selecting the value that matches with the observed colour) as suggested by Harrower and Brewer [2003], Crameri et al. [2020]. However, rainbow colourmaps may be appropriate as more hues allow for easier indication.

As a small-scale system evaluation, the participants of the walkthrough demonstration study only used Stacked. As such, the study was not attempting to fulfill Obj2. Without a baseline technique, we do not make any conclusion on the effectiveness of the technique.

1.1.3 Obj 3: Designing and evaluating an AR+tablet hybrid user interface

An OST-HWD and a tablet are mobile devices that have different capabilities. An OST-HWD can display content in a wider area than a tablet. It can also display 3D dimensionality content. However, it lacks the computational power to compute everything (e.g. fitting a model). A high-end tablet can provide additional computational support. Furthermore, its display has a worse fidelity than a tablet, and its input is less precise [Soares et al., 2021, Feiner and Shamash, 1991]. Combining these two devices into a single hybrid user interface could address each other's shortcomings. The AR helps the user to see beyond the extent of the tablet's display.

Meanwhile, the tablet can display details that the AR cannot and provides more precise input through touch gestures. The tablet can also compute ML models—helping OST-HWD to better manage its limited resources.

Combining two devices can have its own set of challenges. For instance, both devices have different ways of input. An OST-HWD like Microsoft HoloLens v2 supports hand gestures, while a tablet often supports touch gestures. As such, we make it a research objective to understand how to combine both devices into a single hybrid user interface, and how the user reacts to such an interface. In the design chapter (Ch. 3), we discuss the technological implementation that allows the tablet and the OST-HWD to communicate with each other. In the colourmap design chapter (Ch. 4), we provide background information on how the tablet's screen can interfere with AR display and the ways to mitigate the issue. Then, in the synoptic study chapter (Ch. 5), we discuss how the user navigates glyph fields with both devices at the same time. Although the elementary study chapter (Ch. 6) is more focused on glyph comprehension, we still varied the distances of the glyphs from the user to represent the condition where the glyphs are on or off the tablet. In the walkthrough demonstration study, multiple experts had a chance to try out the AR+tablet paradigm.

1.2 Synopses of the Thesis Chapters

We outline the structure of this thesis and a brief summary for each chapter.

Chapter 1: Introduction

In this chapter, we provide the motivation for our work and brief summaries of all other chapters. We outline the research objectives and how our studies fulfill these objectives. Additionally, we briefly state the outcome of our work.

Chapter 2: Related Work and Background Information

We outline related literature from multiple areas such as interactive machine learning (IML), exploratory data analysis (EDA), visual analytics, and more. Finally, we also provide background information necessary for understanding the technical aspects of the thesis. The information includes multiple linear regression (MLR), glyph designs, AR content placement, and types of evaluation. Since colourmaps play an important role in the design of Gander, a large portion of the chapter includes background information on such work. Based on the analysis of the related work, we identified multiple research gaps. The most important ones are: (1) the lack of work in IML with AR+tablet, (2) the lack of empirical evaluation of glyph visualization in AR, and (3) the lack of evaluation of hybrid user interfaces that make use of AR and tablets.

Chapter 3: Exploratory Background Work

Prior to the design of Gander, we performed explorative research to identify the specific areas within immersive analytics that we can focus on. Such research work is necessary due to the transdisciplinary nature of immersive analytics. The work includes a visual query language, a pair of studies [Hu et al., 2021] in out-of-view target acquisitions in virtual reality (VR), and variance statistics. The elements of the work eventually evolved into Gander, and the three studies.

Chapter 4: Design of Gander

We describe the design of Gander, the AR+tablet IML system for this thesis. The chapter does not provide detailed descriptions of the colourmaps used in the system. This chapter touches upon Obj1 and Obj3, as it discusses designing a hybrid user interface for machine learning with geospatial data. It is also relevant to Obj2 as we describe Stacked which is the default glyph-based visualization technique. Although the software used in the synoptic and the elementary studies shared the same codebase with this version of Gander, the software contained significant variations. Please refer to the studies' chapters (Ch. 5, 6) for more information.

Chapter 5: The Synoptic Study–Understanding Glyph Field Navigation

This chapter contains a description of the synoptic study. In this study, the participant interacted with a version of Gander which we modified so all participants completed the tasks with the same sets of variables and models. They performed various tasks relevant to creating and assessing MLR models. We compared Polyline and Radial. We excluded Stacked, because its parallax effect could act as a confounding variable. We analyzed the trajectory data and found that the participants tended to rely on their gaze with Radial, and more on the tablet's touchpad with Polyline. Therefore, we conclude, that shape-based techniques are good for encouraging closer examination while colour-based techniques are good for contextual display. Unfortunately, the participants were not able to provide precise effect sizes. We found that accurately measuring effect sizes is its own highly complex topic.

The study addresses all three research objectives. It measured how the user could perform tasks through the pre-fit and the post-fit stage (Obj1). It analyzed how the user scans the glyph fields (Obj2). Lastly, it involved understanding how the user operates an AR+tablet hybrid user interface (Obj3). The results of the study can be used in the future to help with geospatial software with an AR+tablet interface, and to better understand how the glyphs' appearance affects the use of the hybrid user interface.

Chapter 6: The Elementary Study–Comprehension of Individual Glyph Composites

The glyph field study, while good for understanding scanning behaviour, does not allow us to understand how the user perceives information from each composite. This elementary study supplements by focusing on the perceptual aspect of the glyph. We compared Radial against Stacked to see if the parallax effect used by Stacked can improve the user's understanding of the visualization or reduce comprehension time. The study shows that the parallax effect can be helpful when glyphs are further away; Stacked is faster at a distance due to the glyphs already being decomposed. We also found that the parallax effect has less impact on effectiveness than the complexity of the glyph composite. Essentially, the more glyphs there are in a composite, the slower and less accurate the user becomes. The study also highlights the importance of good colourmaps and visual aids design.

The study mainly focuses on Obj2 as the participants only performed colour-value judgment tasks in the study. However, it also touches on Obj3 when we vary the distances of the composites from the user to simulate the conditions of the glyphs being on and off the tablet.

Chapter 7: The Walkthrough Demonstration Study

We mainly focus on Obj1 for this study. Unlike the previous two studies where we mainly focus on glyph-based visualization, our aim is to understand how experts use Gander as a whole system. We conducted a walkthrough demonstration of the prototype. The walkthrough demonstration has been deployed in some HCI works [Ledo et al., 2018]. We interviewed six people who were involved in spatial analysis in some capacity. Four participants were expert analysts. Two were usability evaluators who worked on MR-based interfaces for geospatial analysis. Five participants interacted with Gander in person while one only viewed the video demonstration. We performed a qualitative analysis similar to the work of Reilly and MacKay [2013] with thick descriptions Nas et al. [2023], Cheung et al. [2014]. We found that the geospatial analysts tended to have flexible practices, so expanding the vertical slice to satisfy all types of users may not be possible. Still, some functionalities (e.g., better tablet interface, summary statistic visualization) are useful to have and are discussed in Ch. 9.

Chapter 8: Discussion

While the study chapters (Ch. 5-7) have their own discussion sections, this chapter bridges them all together. We also re-iterate the three research objectives. Then, we reflect on the design of Gander and the results of the studies based on the research objectives, and the prior literature.

Chapter 9: Proposed Changes to Gander

In this chapter, we provide an update on the design of Gander based on the study results to expand the vertical slice. The update includes a new visualization system that shows aggregated information. Furthermore, it contains new tablet-based interactions to support aggregation. While we initially avoided aggregation, the studies suggest that these are very important. The update is provided as a set of low-fidelity designs.

Chapter 10: Conclusion

We provide the concluding remark of the work. My positionality statement can also be found in this chapter.

Appendices

In Appendix A, we provide the description of two novel techniques for analyzing trajectory data with random walk. These techniques are based on the trajectory analyses performed in the synoptic study (Ch. 5). Since mean-squared displacements (MSD), a trajectory index used in the synoptic study and the elementary study (Ch. 6), is a type of variance, this chapter also contains work on analyzing variance data. Prior to adopting likelihood as an indicator of a model's goodness-of-fit, we performed extensive analysis on variance-based methods such as R^2. However, we found such methods to be insufficiently flexible. As we moved to likelihood-based methods, we found that works on variance analyses no longer fit the body of the main text of this thesis. Nevertheless, the knowledge of dealing with variances (e.g., creating a confidence interval for a variance estimate), is still indirectly applicable to analyses of second-order statistics such as MSDs. Therefore, Appendix A also represents a part of technical work that would otherwise be excluded.

It is important to note that the method proposed in Appendix A is still at a preliminary stage. Trajectory analysis is a complex topic, and a full treatise on this topic is beyond the scope of this thesis. However, by outlining methods in a publicly available document like a thesis, we hope that it may be useful to those who are looking to tackle such a difficult topic.

Appendix B contains copies of the approval letters from research ethics boards. The consent forms can also be found here. We do not add the study instruments in the appendix. Instead, the instrument information is incorporated into the thesis chapters themselves.

Chapter 2

Related Work and Background Information

This chapter provides a discussion of the prior work on various topics that are relevant to the thesis. First, we discuss interactive machine learning (IML). Although Gander is not a full IML system like Orange [Demšar et al., 2013], it still has some functionalities of an IML system. In this section, we also provide background information on multiple linear regression. Then, we provide information on exploratory data analysis (EDA), a type of statistical analysis that plays an important for Gander. Then, we discuss immersive analytics, its evolution, and its relevance to our work. Mixed reality technologies are discussed here. Afterwards, we provide information on colourmap designs for OST-HWDs which require special considerations. Lastly, we discuss using AR in the context of mobile computing and combining AR with another device to create a hybrid user interface.

2.1 Interactive Machine Learning

IML, according to Dudley and Kristensson [2018], is a system that allows the user to be involved through the whole process of fitting machine learning models and analyzing them. An individual can help to curate data, select features, fix issues, and assess the models. This does not mean that an IML needs to exclude all forms of automation though. For instance, we argue that R, a statistical software package, is an IML, because it allows a person to make input to a ML process. However, R also allows for automatic step-wise regression where the machine automatically creates all candidate models and selects the best one [Jenkins-Smith et al., 2021]. An IML system does not need to include all features described by Dudley and Kristensson [2018]. For instance, Gander does not allow for model steering–the user cannot modify the data to achieve better fitting through the system itself. Furthermore, Gander can only perform MLR at this moment. Therefore, when describing the background and technical information on machine learning, we will only provide that which is relevant to MLR.

We cannot consider Gander an IML system due to certain technicalities. In our prototype, we used the ordinary least squared (OLS) method. While OLS can generate predictive models, its most popular use is to create descriptive models. A descriptive model does not make a prediction; rather, it determines if a correlation between the observed data and effects exists [Sheffet, 2017]. However, there are variants of MLR that use iterative gradient descent methods. These variants can be considered as machine learning, because software must iteratively optimize the models using the errors from multiple candidate parameters. In other words, the software "learns" to pick the best parameters [Bottou, 2012]. We can imagine a simple extension to Gander which makes it a true IML system.

2.1.1 Feature Selection

Normally, we want to select the smallest number of features or independent variables (IVs) to fit into the model. This follows the principle of parsimony where we aim to have the simplest model [Jenkins-Smith et al., 2021]. A goal in modelling is to determine the impact of individual features on the data. If our model contains features that are too similar to each other, it can be difficult to establish a correlation. For instance, if we have a model that contains Education Level, and Age as features, and they turn out to be similar, we may not be able to state which of the features affect the response values.

We argue that feature selection is both a manual and an automated operation. In certain cases, the user can easily remove undesirable features early on. An example of this is a researcher designing an experiment. From the outset, the researcher can determine the features that will be included in a model. If they are conducting a study where the user is interacting with a novel interface, they can control various factors such as the machine used, the participant's skill level, and more to ensure that the model is as parsimonious as possible. However, not all users have such luxuries. Some users do not control how the data are produced. For instance, geospatial analysts often have to contend with data already collected. In such cases, IML can be useful. Orange by Demšar et al. [2013], a desktop-based IML with a graphical user interface, allows the user to interactively apply techniques like Principal Component Analysis (PCA) to simplify complex models. Akaike Information Criterion (AIC) and step-wise regression, available through software packages like R, can help us choose variables to exclude based on inference [Snipes and Taylor, 2014, Jenkins-Smith et al., 2021]. Geographic Information Systems (GIS) also can be used for further exploration of features.

2.1.2 Fitting for Multiple Linear Regression

After selecting features, the user wishing to use MLR or related techniques like General Linear Model (GLM) needs to fit the data with the selected features. We can use MLR for a predictive model and a descriptive model. A predictive model aims to predict the values of the dependent variable (DV) based on a set of given feature values. Meanwhile, a descriptive model does not make a prediction. Instead, we create it to assess the strength of the features and how they correlate with the DV [MacKenzie, 2013]. While its name suggests that features must be linearly distributed values, an MLR can also handle non-linear data; we can encode categorical features into numbers, and transform non-linear features into linear ones Peterson and Cavanaugh [2020].

2.1.3 Model Assessment and Model Selection

There are many methods of assessing and comparing models. We considered three methods for Gander: R^2-based method, analysis of variance (ANOVA)-based, and likelihood-based. Ultimately, we chose the likelihood due to its extensibility to more complex models (e.g. GLM).

R^2 is a ratio of two variances: explained and total variance [Lewis-Beck and Skalaban, 1990]. Explained variance is the variance of the model, while the total variance is the variance of the data [Lewis-Beck and Skalaban, 1990]. R^2 is easy to understand. However, they are only applicable to MLR models. For other models, we can use similar effect sizes called pseudo-R^2. While pseudo-R^2's resemble R^2, they can be computed quite differently. Some pseudo-R^2, like Nakagawa's R^2_{GLMM} [Nakagawa et al., 2017] uses different types of variances. However, some rely on other types of estimates such as likelihood [Nagelkerke, 1991]. Regardless of how R^2 was obtained, to use it for model selection, the user generates multiple candidate models and selects the one with the highest R^2.

An ANOVA-based method involves comparing two types of variances: the variances explained by the "full model", and the "restricted model" [Glatting et al., 2007]. The full model is one with more variables. We abandoned the ANOVA-based method, because it is unclear how sums of squared can be computed for models beyond MLR. Unlike R^2, we choose models based on its p-values. If an ANOVA found two models to be sufficiently different, then the p-value would be low. Otherwise, high p-values indicate the models are not different from each other.

Although we have already mentioned using likelihoods for computing pseudo-R^2, its flexibility lies beyond to usage as effect sizes. For example, Akaike Information Criterion also uses likelihood for inferences [Snipes and Taylor, 2014]. Such flexibility leads us to adopt likelihoods for Gander. A likelihood function represents a probability of a set of estimates being the true parameters given the observed data. First, we compute $f(x_i|\theta)$ where f is the probability of x_i being the DV values if we use parameters θ in the model. Based on the type of ML, the function f can vary. Below are the examples from [Isoni, 2016]:

- Multiple Linear Regression: $f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\epsilon_i^2}{2\sigma^2}}$ where ϵ_i is an error and σ is the standard deviation of all errors.

- Logistics Regression: $f(x_i|\theta) = (\hat{y_i})^{y_i}(1 - \hat{y_i})^{1-y_i}$.

The values inside θ can vary based on the type of ML. For instance, in MLR, θ is a set of the fitted coefficients. We assume that $L(\theta|x_i) = f(x_i|\theta)$ and the likelihood of the model, given all x is:

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta).$$

If we have multiple candidate models, we can use Maximum Likelihood Estimate (MLE) to select the best model. In this case, we can simply pick the model with the maximum L. It is important to note that the candidates must be nested; their θ must be subsets of the full model's θ. The full model is the model with all IVs which makes it the least parsimonious model possible. If we also want to penalize models for not being parsimonious, we can compute Akaike Information Criterion (AIC) values from L. AIC, unlike MLE, includes a penalty for having too many coefficients [Snipes and Taylor, 2014].

2.1.4 Multiplicativity

Sometimes, a single IV does not affect the DV as much as when the feature is combined with other ones. We can call this an interaction effect–or multiplicativity. We can represent multiplicativity in a multiple linear regression (MLR) model as a separate term. According to Friedrich [1982], Braumoeller [2004], multiplicativity can be difficult to understand.

2.1.5 Spatial Autocorrelation

In geospatial analysis, the positions of the data and their closeness to other data points can affect their own properties. For instance, we can expect the data from Halifax, Canada to resemble each other more than data from outside of the city. Gander allows the user to find if there could be any regional difference in data and goodness-of-fit. Tobler once stated [Miller, 2004]: "I invoked the first law of geography: everything is related to everything else, but near things are more related than distant things." While Miller [2004] sees this statement as a good guiding statement in geography, he argues that we also must have other measures of similarity. He states, it is possible that "near things" are similar to each other simply due to coincidence.

Spatial autocorrelation is one of the measures that indicates if the data are geospatially clustered. Some examples of autocorrelation are Moran's I, Getis-Ord G statistics, Local Indicators of Spatial Autocorrelation (LISA), and Getis-Ord G statistics [Zhang and Tripathi, 2018]. Zhang and Tripathi [2018] computed these types of autocorrelations to study the impact of PM2.5 (Particulate Matter 2.5) pollution in Eastern Thailand on lung cancer. They state that while Moran's I is popular, it is inappropriate if the data contain hot spots or a grouping of patterns. To deal with the hot spots, they have to use LISA to locate the hotspots and then use Getis-Ord G statistics to compare their shapes. According to Zhang et al. [2009], there are three ways that we can incorporate spatial information into our own regression models: (1) space lag model (SLM), spatial error model (SEM), and geographically weighted regression (GWR). SLM assumes that there is a lag caused by spatial association. SEM assumes that the spatial information only affects the error terms regression. Meanwhile, GWR generates local models based on local weights [Comber et al., 2023]. According to Comber et al. [2023], properly applying GWR requires generating multiple candidate models and user input. Since selecting a correct autocorrelation strategy can be difficult [Getis, 2010], visualizing data can be helpful. For example, without visualization, we may not realize that our data contain hot spots.

Peña-Araya et al. [2020] developed map-based 2D visualization techniques that vary based on geolocation and temporal information. They compared the techniques through a comparative study where the participants tried to identify correlations between the data and the spatial and temporal data. The researchers found that using small multiples, including glyphs, was more helpful for the user to detect correlation between space and time.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a way to better understand the data. Hoaglin [2003] states that early EDA activities involve: (1) identifying outliers, (2) smoothing the data, (3) transforming data into distributions that are easier to work with, and (4) data visualization. Initially, due to the lack of computational power, the early visualization techniques were limited to techniques such as stem-and-leaf plots, and box-and-whisker plots. Modern software, like Tableau [2016], is more capable and can generate additional types of visualizations (e.g. pie chart, heatmaps) on the fly. Gander does not focus on the four original tasks identified by Hoaglin. Instead, it focuses on providing visualization to assist the user in geospatial analysis. Still, Gander still follows the spirit of EDA by using visualization and data exploration to support the user in preparing the data for analysis.

The work by Andrienko and Andrienko [2005] discusses applying EDA to trajectory analysis. Although we were unable to apply their methods to our trajectory data in the synoptic and elementary studies due to random walk (i.e. any random movements–not necessarily restricted to walking), their concepts of synoptic and elementary tasks still helped us to frame our study. Andrienko and Andrienko [2005] state that visual analytics tasks can be separated into two types: elementary and synoptic. An *elementary task* involves the user finding a single simple piece of information. A study by Jankun-Kelly et al. [2010] is an example of this; in this study, the participants identified the individual values expressed by one 3D glyph per trial. A *synoptic task* involves identifying information that requires aggregation of other information. For instance, in Yang et al. [2022]'s study, the participants identified population density in five contiguous states of the United States of America. The participants were interacting with a VR visualization system that was altered based on the VR controller's orientation. These works guided the designs of the synoptic or elementary studies (Ch. 5, 6).

EDA can be useful with map-based visualization, because map-based visualization involves a large degree of exploration. For instance, if a user tries to learn about gas stations along a highway, they can look at the highway to identify all gas stations. They can also zoom and pan the map to gain more insight. It is important to note that EDA does not require this type of exploration. For example, the user could also perform an EDA by typing in a search query, and a system simply outputs all gas stations in a text list. On a map application like Gander, the user can follow Shneiderman [1996]'s paradigm: looking around to gain overview knowledge, zooming onto the map to specific areas, filtering out irrelevant information, and then obtaining details. All actions should be done on-demand—i.e. whenever the user wishes to. In the gas station example, the user starts with a low zoom setting to gain insight into the general area. Then, they zoom into a specific part of the highway. Finally, they query for information about specific gas stations.

Schneiderman's paradigm is not the only one for interacting with map-based visualizations. Baudisch et al. [2003] mention other possible paradigms. For example, *overview+detail* (O+D) has two simultaneous visualizations: the overview map showing the larger area, and a more detailed map. Another is *focus+context* (F+C), where two display resolutions are used. One, with a higher resolution, is where the user focuses on and gleans information. Another one, with a lower resolution, is outside the user's focus area and is to provide contextual information [Baudisch et al., 2003].

2.3 Immersive Analytics

Dwyer et al. [2018] define immersive analytics as "the use of engaging, embodied analysis tools to support data understanding and decision making." The field involves the use of MR to enhance the user's immersive experience. Ens et al. [2021] argue that designing studies for immersive analytics can be difficult due to the novelty of the field. However, the existing study designs from information visualization and visual analytics, a precursor to immersive analytics Dwyer et al. [2018], can serve as useful blueprints. In this section, we review works from these areas plus Digital Earth and glyph-based visualization. Digital Earth is a topic pertaining to the use of immersive analytics for geospatial analyses [Çöltekin et al., 2020]. Meanwhile, glyph-based visualization is a technique that has been adopted for general types of data visualizations–including immersive ones (e.g. Lau et al. [2019]), and geospatial analyses (e.g. Peña-Araya et al. [2020]).

2.3.1 Information Visualization

Information visualization predates the field of computer science. Some visualization techniques, like the pie chart [Funkhouser and Walker, 1935, Spence, 2005], have existed before the creation of the computer. One of the oldest canonical examples is the visualization created by Snow in 1854 [Brody et al., 2000]. Snow developed a map-based visualization of a cholera outbreak in 1854 in order to communicate his theory positing the outbreak was caused by a contaminated water pump. He used small multiples of bar charts. This example, alongside Snow's contemporary map-makers, represents an early attempt to incorporate maps into visualization. Contemporary examples of this include the work by Ondrejka [2016] which uses modified cartograms to represent proportional election information. Another one by Kumar et al. [2013] overlays pixels over maps to represent predicted data. The design of Gander is partially inspired by these past and contemporary works.

2.3.2 Visual Analytics

Visual analytics is an extension to information visualization [Dwyer et al., 2018] which includes interaction. Instead of having a static visualization, the user can easily and actively modify the visualization. According to Perin et al. [2014], Bertin's Matrices represent an early attempt at visual analytics. The user operates a physical table that they can modify to represent table-based data. Perin et al. [2014] themselves developed a computer-based adaptation of the table. For map-based visual analytics, Willett et al. [2015] produce a prototype where the user can better understand terrain heights. The tablet display allows the user to "jiggle" the terrain, and the "jiggling" animation is affected by the heights of the terrain themselves. Gander includes some modes of interaction. For example, the user can change the features and the models being displayed on the system.

2.3.3 3D Visualization and Immersion

Before we can get to immersive analytics, we must first discuss work in 3D visualization. We, humans, are creatures living inside the realm with three physical dimensions. While 2D visualizations are useful, we do not feel embodied within them; instead, we feel detached. Brath [2014] argues that we have an innate ability to understand 3D visualization. Since we are living in a 3D physical world, we have instinctive abilities to understand 3D structures. This does not mean that 3D visualization does not have any pitfalls [Brath, 2014]. Szafir [2018] argues that certain types of charts like 3D pie charts, can distort information and mislead people.

3D visualization is important for creating a sense of immersion–a vital element of immersive analytics. With OST-HWDs, we can implement large-scale visualizations the user feels immersed in. The visualizations themselves can also become part of the physical world as situated visualizations [Bressa et al., 2022]. Data physicalization can further induce a sense of immersion through the use of physical objects to represent information [Hull and Willett, 2017]. However, AR is still more convenient in the sense that we can easily generate 3D objects without physical objects (e.g. OST-HWD or a cellphone).

2.3.4 Digital Earth

Digital Earth is a digital twin of Earth that the user can use to analyze its geospatial properties [Çöltekin et al., 2020]. Jones et al. [2020] define a *digital twin* as: "A complete virtual description of a physical product that is accurate to both micro and macro level." As a virtual copy, the user can interact with it in any way they wish. For instance, they may destroy parts of the copy to better understand the Earth's inner properties. This copy of Earth can come in various fidelities. For instance, if the user only requires overview information, they simply need to view it through a 2D interactive map application. An example of this is by Kumar et al. [2013]. They created a predictive linear regression model for biomass on a map. To visualize the output, a layer of pixels representing the predicted values were superimposed onto

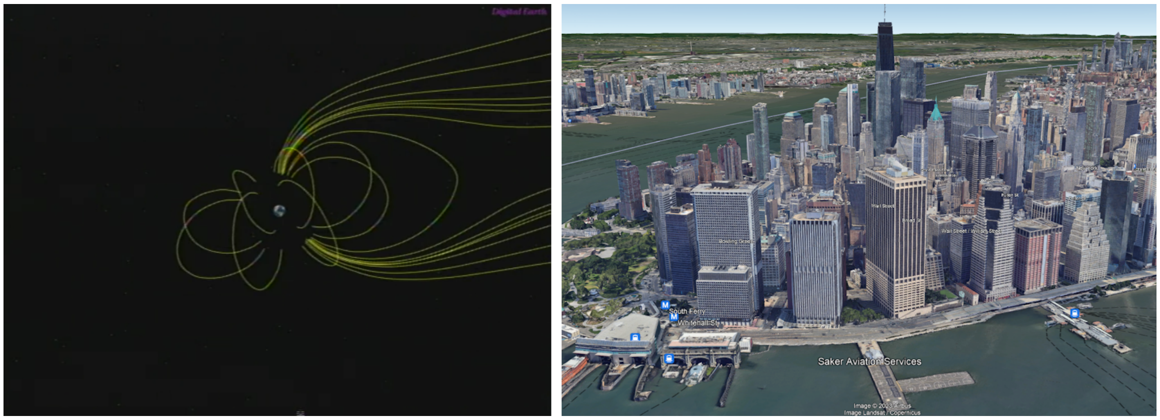


Figure 2.1: Examples of software relevant to the Digital Earth. **LEFT:** NASA Digital Earth Workbench, VR software, showing the Earth and its magnetic field. The screenshot is from [Maher and Spicer, 1999]. **RIGHT:** Google Earth (Desktop Version) [Google, 2023] showing the 3D models of building in New York City.

the map. Gander, with its glyph-based visualization, took inspiration from this work.

If the user wants an immersive experience where they feel embodied, they will need the use of mixed reality. The use case dictates the degree of immersion required. For example, if a user requires 3D visualization, but does not need full isolation from the physical world, the software can be similar to Airbus Holographic Tactical Sandbox. The Sandbox is a war-planning tool that requires the use of OST-HWDs. It provides an overview of a geographic area and helps the user to plan combat activities [Walsh et al., 2023]. Another example is Tangible Globes [Satriadi et al., 2022] where the user can use a spherical input device to better understand geospatial information. We further discuss the evaluation of this device in Section 2.6. Gander is aiming for this level of fidelity; the user has an overview of the Digital Earth, and is not embodied in it.

2.3.5 Glyph-based Visualization

Glyphs are visual markers that we place onto locations to present information in those specific locations. Due to the versatility of glyph-based visualization which allows it to be used in information visualization, visual analytics, and immersive analytics, we dedicate a portion of our literature review specifically to glyph-based visualization.

According to Borgo et al. [2013], MacEachren et al. [2012], we can change a glyph's appearance through multiple visual channels, such as colour, shape, size, orientation, texture, and more. For instance, Gander uses colour to denote a value in the pre-fit stage. In this scenario, the user must transform the colour into the value through their cognitive processes. The work by Borgo et al. [2013], MacEachren et al. [2012] is influenced by visual semiotics, a branch of semiotics that is concerned with how we glean information and obtain meanings from visual media MacEachren et al. [2012].

There are myriad ways to place a glyph–i.e. affecting the glyph's position visual channel. In Gander, we use transformed longitude and latitude coordinates or the Cartesian coordinates to determine the locations of the glyphs. If there is no spatial data available, we can use spatialization to assign the positions of the glyphs. Spatialization is a process of assigning spatial information using non-spatial data [Skupin and Fabrikant, 2003]. For example, Lau et al. [2019] created an AR-based immersive visualization that represents genomic data as glyphs. The glyphs that represent more similar genomic data are placed closer together. Lau et al. [2019] obtained qualitative feedback from five participants; they did not conduct any empirical user study to determine its actual usability.

In some visualizations, a glyph needs to present more than one variable. To do so, we target multiple visual channels. For instance, if we want to represent velocity at a point, a glyph will need to not only present the speed but also the direction. In such a case, we could use an arrow as a glyph with the direction of the arrow being the same as the direction that we want to represent, and the length (or size) of the arrow being the speed. Another example is Z-Glyph created by Cao et al. [2018]. A Z-Glyph is a n-pointed star with each point representing a variable. Inside the star, there is a circle that serves as a baseline value. A point extends out far away from the circle if the value's Z-value (i.e. Z-statistic) is quite high. If the Z-value is very low (further away from zero), then it points inward instead. The user can use the positions of the points relative to the circle to detect outliers. An alternative is to create a composite glyph which is a combination of multiple glyphs [Ropinski et al., 2011]. According to Ropinski et al. [2011], glyph composites can convey more complex information than a single glyph. An example of this is a Chernoff face. A Chernoff face represents a facial expression. Each Chernoff face has separate glyphs that can be independently manipulated [Fuchs et al., 2017]. Fig. 2.2 shows examples of the glyphs.

Some glyphs are small multiples–i.e. each glyph is a small chart based on existing techniques [van den Elzen and van Wijk, 2013]. The visualization by Snow [Brody

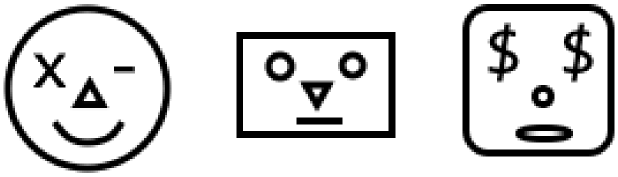


Figure 2.2: Examples of Chernoff faces. The shapes in each face can be independently manipulated–making the face a glyph composite.

et al., 2000] is an example based on bar charts. Another bar chart-based example can be found in the work of Peña-Araya et al. [2020]. In their work, they used small bar charts whose x-axis represents time and y-axis represent heights. They compared the bar chart technique with two alternatives: the Dorling cartogram, and proportional circles (i.e. circles whose sizes varied based on the values). These techniques did not convey temporal information. They found that the temporal information shown in the bar charts was helpful. Sun and Kuo [2002] used small multiples of monochrome heatmaps to represent trends in matrices. They did not evaluate their technique. While Sun and Kuo [2002] used a different type of spatialization, we note that their technique was somewhat similar to Radial and Stacked. Originally, we planned for our Radial and Stacked glyphs to be monochrome. However, due to most OST-HWDs lacking a subtractive display, we had to implement non-monochrome colourmaps. Otherwise, dark glyphs would appear invisible to the user. The final example is Polyline by Opach et al. [2018] which is based on line charts. The evaluations of the technique [Opach et al., 2018, Opach and Rød, 2018] show that Polyline is good for understanding values as opposed to detecting differences between glyphs. Because of this, we used Polyline as the control technique in our synoptic study; our synoptic study involved understanding the glyphs' values.

Glyph-based visualization provides a more detailed view of the data. Therefore, it can help the user to avoid the *ecological fallacy*–where aggregated information like aggregated information is assigned to all data points [Salkeld and Antolin, 2020]. However, glyph-based information can lead to the opposite type of fallacy. Display granular data can risk the user conflating locally observed data to be a global trend [Zhao et al., 2017]. For instance, after seeing a few glyphs indicating low values, the user concludes that the average for all values must be low–even though they need to see more glyphs to be certain. This type of error is called the *atomic fallacy* [Keskin, 2022]. Furthermore, glyph-based visualization can suffer from other issues such as overdrawing and visual clutter.

Even though Gander can only perform glyph-based visualization, it does not suggest that we are eschewing other types of visualizations. Rather, our main interest is to understand how the user gleans information from individual data point, and aggregate them. In our thinking, once we have developed a user model for understanding invidual data, we can then create the visualization method for aggregated information that meshes well with the user model.

Overdrawing and Visual Clutter

Overdrawing is an important issue in glyph-based visualization. When glyphs are too close to each other, they can impact the overall comprehensibility of the glyphs [Mayorga and Gleicher, 2013]. Visual clutter is when the visual aspects of the environment start to interfere with the comprehensibility of the interface [Rosenholtz et al., 2007]. Since OST-HWDs have translucent screens, the physical world itself can impact the interface. While Gander does not have any mechanism to prevent overdrawing and visual cluttering, these concepts influenced how we chose our data, and presented them. The maps that we chose for our studies tended to have data points that were relatively spread out. Furthermore, we chose the zoom level that reduces overdrawing. Although, in theory, we can eliminate all overdrawing by using a large zooming scale, the glyph-field navigation can become frustrating for the user since they have to scan a large distance.

Apart from zooming, there are other methods of dealing with overdrawing and visual clutter. First, we can simplify the overdrawn area itself. Mayorga and Gleicher [2013] propose a visualization technique called Splatterplot as an alternative to scatterplot. Instead of drawing all glyphs in a scatterplot, Splatterplot renders some of the glyphs as examples. Contours or large shapes then replace the remaining glyphs. If the chart presents multiple variables, then Splatterplot renders contours of different colours. Where the contours overlap, the plot can blend their colour differently to convey relational information between the variables. Mayorga and Gleicher [2013] never evaluated their techniques. However, they applied the technique beyond the scatterplot. In one of their case studies, they visualized fatal car crashes in the United States of America (USA). They found that contours exist in populated areas. In sparsely populated areas, the data are still presented as glyphs.

McNabb and Laramee [2019] propose that we can add interaction to resolve overdrawing issues. For instance, Gander allows the user to select the features and the models that are drawn on the screen. This allows the user to ignore the glyphs that they do not want to analyze at the moment. Another possible functionality is zooming in and out. However, zooming can be a complex issue. While zooming in can resolve overdrawing by adding distances between glyphs, zooming out can introduce the issue. Furthermore, zooming in and out can affect the number of glyphs that user can see, which affects how they make inferences. This requires many considerations. Since our work is more focused on glyph comprehension, and navigation and less on other types of interaction, we keep the zoom level static.

Overdrawing is related to a concept in vision science and AR called visual clutter. According to Rosenholtz et al. [2007], clutter can occur with any sense–not just vision. They define clutter as: "Clutter is the state in which excess items, or their representation or organization, lead to a degradation of performance at some task." There are some visualization and AR studies that aim to analyze how we can deal with clutter. On the other hand, in AR, Peterson et al. [2009] evaluated three virtual label-placing techniques to reduce visual clutter. In the first technique, depth-based, the labels have different stereoscopic disparity to represent different depths. In the second technique, height-based, the labels have different heights, so they do not occlude each other. In the last technique, planar, the labels can be separated in many directions on a plane. They found that all separation techniques were faster than the control one; however, they could not pinpoint the most effective techniques. We note that Stacked, one of our glyph composite techniques, can be optionally separated. When the user views Stacked glyphs from certain angles, the glyphs could be totally separated due to the parallax effect (see Rouan [2015] for more information on the effect).

The work by Hu et al. [2021] suggests that the physical world itself may serve as a source of clutter. In their VR study, the participants used a visual cueing technique to reach out-of-view targets. One of the conditions was the environment with varying degrees of visual clutter. They found that the environment could interfere with the user's performance. In addition to the visual clutter, Ferrer et al. [2013]'s work on phone-based AR demonstrates that background motion in the physical world can also worsen user performance. Since Gander relies on two mobile devices (an OST-HWD and a tablet) for display, the user can, in theory, deploy the software anywhere. Therefore, Gander itself relies on the user selecting a good environment with minimal visual artifacts.

Evaluating Glyphs

There are multiple ways of evaluating glyph-based visualization. Fuchs et al. [2017] argue that glyph evaluations could be categorized into two major types based on the work of Andrienko and Andrienko [2005]: synoptic and elementary. A synoptic study contains three types of tasks which include: (1) searching for specific glyphs, (2) looking for similarities between groups of glyphs, and (3) finding trends and correlations among the glyphs [Fuchs et al., 2017]. Our synoptic study contains the elements of these tasks. The participants must look for the minimum and the maximum glyphs, they must check similarities of multiple variables/models, and they must find trends which include spatial autocorrelation, and correlation. According to Fuchs et al. [2017], an elementary study contains simple lookup tasks, or a task that focuses on a single piece of information. Our elementary study's task follows this by asking the participants to only indicating various values in a composite. Fuchs et al. [2017] state that there are fewer elementary studies, because naturally, a user relies on multiple glyphs to make decisions.

Ward [2008] argue the research questions for glyph-based visualizations are as follows:

1. How does the user rank the glyphs based on their attributes (e.g., colour)?
2. How quickly and correctly does the user perceive glyphs?
3. How does the user understand glyphs when there is a visual clutter (i.e. glyphs overdrawing)?
4. How well does the visualization technique scale from displaying a small number of glyphs at a time to showing a glyph field?

Our glyph evaluation studies (i.e. the synoptic and the elementary studies) contain elements of 2-4. In both studies, we attempted to measure accuracy. In the synoptic study, the glyph field served as a clutter and the participants must use scanning and navigation to resolve the clutter. Both studies represent glyph evaluation at different scales (a single glyph composite per trial v. a full glyph field).

Siva et al. [2012] argues the general tasks for glyph evaluations are:

1. Identifying data within a glyph: What is a value represented by a glyph?
2. Comparing data: What are the trends displayed by the glyphs?
3. Assessing overall state of the system: What are the global trends displayed by the glyphs? Unlike "comparing data", this task focuses on the gist–or the general information of the whole scene as opposed to localized trends [Sampanes et al., 2008].
4. Monitoring glyph dynamics: If the glyphs can change overtime, what changes can be detected?

We note that Siva et al. [2012]'s categorization contains elements of longitudinal experimentation, and are not as applicable to our study designs. Siva et al. [2012]'s tasks are also difficult to quantify, and qualitative analysis may be necessary. Therefore, they are useful for studies that have higher external validity. Our glyph evaluation studies, on the other hand, are designed to be more experimentally controlled.

Glyphs in Augmented Reality

Prior work applying glyph-based visualizations in AR did not focus on evaluating the glyphs themselves. Instead, AR visualization prototypes are designed and evaluated holistically. For example, Lau et al. [2019] developed an AR visualization system that transforms genomic data into human-shaped glyphs and evaluated the system by simply applying their technique to a dataset without any empirical evaluation. MARVisT by Chen et al. [2020] allows a user to create AR visualizations using glyphs shaped like real-world objects (e.g., bank notes, cars). Chen et al. evaluated the system with experts and non-experts, focusing on user experience and not on the effectiveness (e.g., accuracy) of the glyphs. Büschel et al. [2019] explored 3D renderings of graph networks in AR. The graph networks had two components: (1) nodes, which were represented using spherical glyphs, and (2) edges, which were represented using a variety of lines. The participants indicated if there was any connection between one node to another. They found that the appearances did not have much effect on accuracy. Since Büschel et al. only varied the appearance of the edges, the participants did not perform any glyph interpretation task. Therefore, their study did not reveal any new insight into glyph comprehension.

Alternatives to Glyphs

In addition to glyph-based visualization, we also explore other types of visualizations, because our goal is to eventually design a visualization that can smoothly transitions from displaying indvidual data to representing aggregated data. Two examples of techniques that use aggregation are: density plots, and cartograms. The former involves aggregate local data to prevent overdrawing, and the latter fully aggregate data of a full area.

An example of a density plot is the work by Kumar et al. [2013] who density plots based on linear equations to predict forest biomass in Sariska Tiger Reserve, India. The researchers used satellite images of the reserve to compute the biomass

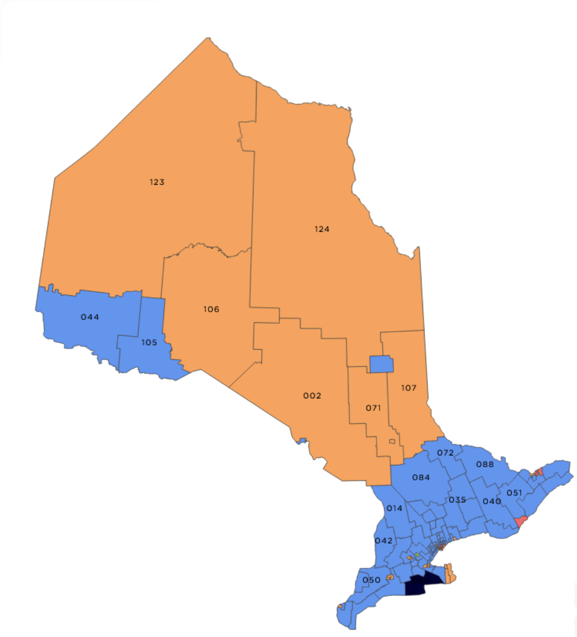


Figure 2.3: Ontario 2022 election map. Orange represents districts won by the New Democratic Party of Ontario. Light blue represents districts won by the Progressive Conservative Party of Ontario. Red represents won by the Liberal Party of Ontario. Dark blue means the district is won by an independent politician. The graphic is modified from Elections Ontario [2022].

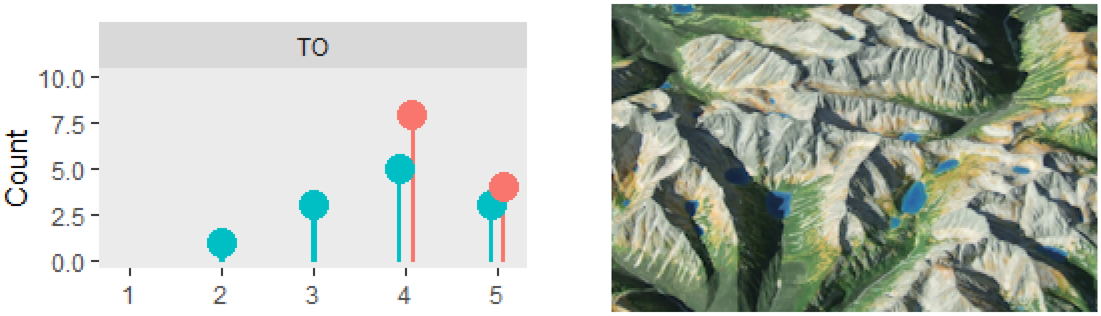


Figure 2.4: **LEFT:** An example of a lollipop chart. **RIGHT:** An example of handmade terrain raised-relief map; modified from Zasdani [2007].

values, and then plotted the predicted values back onto the original maps. Although we considered a density plot as a candidate technique, we did not proceed with the technique because Gander's goal is to display information at the level of individual data points. A density plot meanwhile, can blend neighbouring information together.

2.4 Colourmap Design for Optical See-through Head-worn Displays

The cartogram involves colouring or texturing different sections of a map (e.g. provinces, states, city boundaries) to convey aggregated information of that area. Fig. 2.3 is an example of cartograms representing the results of Ontario 2022 provincial election. Since the sizes of cartograms are based on geographic boundaries, their sizes can skew comprehension of their values [Duncan et al., 2021]. This can be observed in Fig. 2.3; the cartograms show that the New Democratic Party of Ontario has a larger land coverage. However, the sizes of the districts won do not determine the winner of the election. Rather, the number of districts do. As it turns out, the Progressive Conservative Party won in more districts, but their districts tend to be smaller. To mitigate the biases, we can modify the shapes of the cartograms. For instance, boundaries could be distorted so that the cartograms' sizes better reflect the value [Duncan et al., 2021]. Duncan et al. [2021] conducted a study to determine if allowing the user to interactively alter the size can improve inference. They found that interaction is helpful for synoptic tasks (e.g., summarizing pieces of information from multiple cartograms). However, for elementary tasks (e.g. finding the maximum value) adding interaction is not helpful because cartogram sizes do not affect elementary tasks. An alternative to altering the cartogram sizes is to change the opacities of the cartograms. This technique is called *value-by-alpha mapping* by Roth et al. [2010]. However, Peichao Gao and Qin [2019] found that value-by-alpha mapping was less effective than bivariate area cartograms which are cartograms that have both their sizes and colours altered.

Prior adopting glyph-based visualization as the primary method of showing individual data, we briefly considered lollipop plots (e.g., Fig. 2.4-LEFT) and terrain-based visualization (e.g., Fig. 2.4-RIGHT). A lollipop is essentially a modified bar chart; the bar's width has been reduced and there is now a shape at the top [Din, 2019]. Like a bar chart, the height still determines the value. We abandoned this approach due to the difficulties of extending the techniques to represent multivariate information. Furthermore, the terrain-based visualization can be confused with the map terrain itself.

A colourmap (a.k.a, colour scheme [Harrower and Brewer, 2003], colour ramp [ArcGIS, 2022]) is a colour scale that can represent values of either categorical or numerical scaling types [Harrower and Brewer, 2003]. Gander primarily uses colourmaps to convey values in its visualization. Additionally, we found applying the colourmap design principles from information visualization to immersive technologies, particularly the ones like Microsoft HoloLens, is extremely challenging. The lack of literature in designing colourmaps with immersive technologies also compound the issues Erickson et al. [2020]. Therefore, in this section, we provide extensive background information based on colourmap design.

2.4.1 Optical See-through Head-worn Display and Additive Display

OST-HWDs (e.g., Microsoft HoloLens) and video see-through HWDs (VST-HWDs) are devices capable of immersive displays. Both aim to make the user feels surrounded by virtual content. The screens of OST-HWDs allow the light of the physical world to go through the screen. Most of the current OST-HWDs use additive screens to display the virtual content to the user [Kress and Cummings, 2017]. An *additive display*, unlike a desktop display that we are more familiar with, works by modifying the light of the physical world so that it matches the colours of the virtual objects

[Kim et al., 2019]. While the OST-HWDs' additive screens distort the visual images [Itoh et al., 2021], OST-HWDs are more comfortable to use than VST-HWDS because of a higher field-of-view (FOV), and lower latency [Freiwald et al., 2018]. According to Freiwald et al. [2018], higher latency induces more cybersickness.

The opposite of an additive display would be a subtractive display. A *subtractive display* block light instead of allowing it to go through which allows darkness to be represented [Itoh et al., 2021]. *segmented dimming* is a specific type of subtractive display which only subtracts the light for specific objects instead of the whole screen [Magic Leap, Inc., 2022]. In 2022, Magic Leap, Inc. [2022] announces Magic Leap v2 which is the very first OST-HWD that is capable of segmented dimming.

2.4.2 Colourmap

A colourmap is a sequence of colours that represent a set of values. According to Harrower and Brewer [2003], there are three types of colourmaps: qualitative, sequential, and diverging. A qualitative colourmap represents nominal values or categories; examples of nominal values in geospatial data include: the type of the building, the name of the district, languages spoken in a dwelling, and more. A sequential or a diverging colourmap, on the other hand, represents bounded numbers–i.e. ranges of numbers with clear minimums and maximums. They can represent values with ordinal, interval, and ratio scaling (see MacKenzie [2013] for more information). According to Harrower and Brewer [2003], a diverging colourmap has a separate colour representing values in the mid-range while a sequential one does not. Ideally, colours used in sequential or diverging colourmaps are sortable [Crameri et al., 2020]. When viewing two objects with different values from a colourmap, they should be value-orderable based on their colour alone. It is important to note that much of colourmap literature often lacks empirical evidence on their effectiveness in value-judgment tasks–particularly for OST-HWDs [Erickson et al., 2020]. Several studies [Quinan et al., 2017, Reda and Szafir, 2021] found that the rainbow colourmaps may outperform divergent ones. Meanwhile, Gołębiowska and Çöltekin [2022] argue that tasks themselves should determine if the rainbow colourmap should be used or not.

Brightness, Luminance, and Lightness

Light plays an important role with OST-HWDs; unlike traditional screens where the light from the physical world is not a major factor, light can negatively impact the performance of the devices [Kruijff et al., 2010]. Logvinenko [2005] states that perceived brightness can be different from the actual amount of photons in the environment. He defines this quantity as "lightness." On the other hand, "luminance" is a mathematical description of light [Salomon, 2011, Schreuder, 2008]. A colourmap that maintains the same luminance regardless of the value is isoluminant [Kovesi, 2015]. While an isoluminant colourmap is not sortable for those who suffer from strong colour-vision deficiency (CVD), Kovesi [Kovesi, 2019] states that such a colourmap is still useful for scenarios where luminance is used to represent other information. For instance, relief shading of a terrain map uses luminance to describe terrain heights; if the colourmap is not isoluminant, then it can clash with the terrain information [Kovesi, 2019].

Brightness, lightness, and luminance can be different–even if the apparent colour is the same. For instance, a dark grey shadow displayed on an LCD screen may contain more photon than a natural shadow with the same colour. Furthermore, when designing a mixed reality interface for an OST-HWD, we should make sure that the background elements (e.g. a panel of a virtual signboard) are darker than the content [Park et al., 2021, Livingston et al., 2009, Erickson et al., 2021]; Erickson et al. [2021] conducted a study which demonstrates that when the virtual background is dark, the visual acuity is higher and the interface is more comfortable for the user to use.

2.4.3 Criteria for Colourmap Design

We developed the following criteria based on additional research on OST-HWDs with additive displays, and our experience implementing the colourmaps for the device.

C1: Device-based Distortion

As noted by Itoh and Klinker [2015], OST-HWDs have colour distortion stemming from at least two factors. The first factor is the hardware itself can distort the light of the physical world. For instance, Microsoft HoloLens has tinted screens that darken the user's field of view. The second one, which Itoh et al. [2021] also point out, is that the curved screen distorts the user's perspective. How the curvature of the screen distorts the colour is also very dependent on the viewing angle. To see the most accurate colour, the wearer of a Microsoft HoloLens must: (1) adjust the headset so the screen aligns with the user's field-of-view, (2) align their foveal vision to the centre of the screen, and (3) disregard colours in their peripheral vision.

C2: Darkness Is Transparency

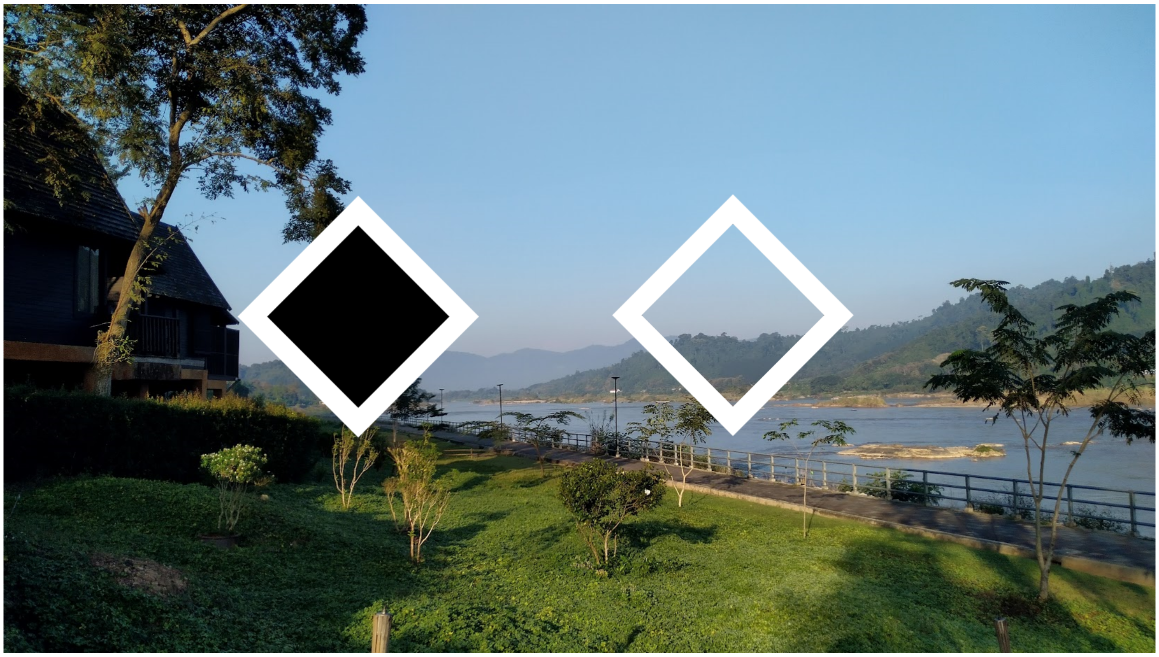


Figure 2.5: The left diamond represents a black diamond shape with a white outline displayed that one may expect. The right diamond shows how OST-HWDs render it; the black fill is completely transparent.

OST-HWDs without subtractive display (e.g., Microsoft HoloLens) treat dark objects as translucent (See Fig. 2.5). For example, a black object will appear completely invisible to the user. This means that if our colourmaps contain colours with lower luminance, the objects with lower luminance will become less visible to the user. Therefore, the objects with higher luminance can appear overly emphasized. As such, we suggest the use of an isoluminant colourmap. Kovesi [2015] defines an *isoluminant* colourmap as maintaining the same luminance, regardless of values. An isoluminant colourmap can be less accessible to those with CVD, because it is less sortable than the non-isoluminant ones such as Viridis [Nuñez et al., 2018, Crameri et al., 2020]. In the future, this issue will be alleviated with the adaptation of subtractive display.

C3: Mind Your Mount–Bright on Dark

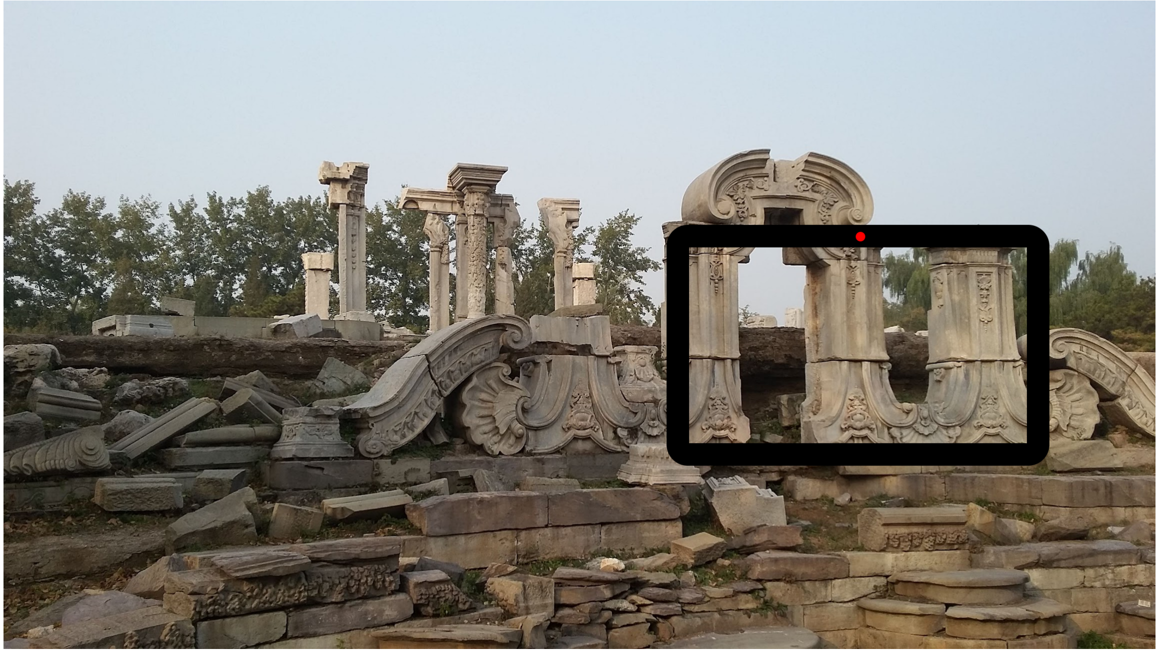


Figure 2.6: While a display screen may have the same lightness with the environment, the actual luminance can be much higher. The OST-HWD may have difficulty mounting on such area.

Tönnis et al. [2013] define mounting as placing virtual content on top of a physical object. When mounting a virtual object, may we suggest: "bright on dark." The object should be brighter than its physical target due to OST-HWDs being particularly sensitive to background light. If the environment is too bright, they may no longer be able to display the content correctly. To demonstrate why mounting a bright object can be problematic, let's assume that we have a very basic OST-HWD that does not have any black tinting on its screen or any correction mechanism. If we want to display blue with the RGB value of $RGB : (0, 0, 0.5)$[1] and the physical world is completely dark, the HWD will simply add 0.5 to the blue channel. However, if the physical world is too bright, the device will not be able to darken the light to the desired colour. While OST-HWDs use different strategies to deal with ambient luminance, their methods are imperfect Itoh et al. [2021]. Itoh et al. Itoh et al. [2021] argue that in order for an OST-HWD to be able to fix any ambient light, it cannot

[1] For RGB, we use the value between 0 and 1, instead of 0 and 255 to keep it consistent with the shader pseudocode.

solely use an additive display. The display must also support subtracting light. We should be especially careful with a light-emitting target like a tablet's screen. Fig. 2.6 shows that while a tablet may have the same lightness as the physical world, the tablet might actually be brighter than the world itself.

Segmented dimming can help to alleviate this issue by blocking the light from the physical world. As such, devices like Magic Leap v2 will fare better than Microsoft HoloLens v2. However, the dimming capability might not be perfect—especially, when there is too much light. Therefore, we must still proceed with caution.

Our "bright on dark" suggestion also applies when mounting a virtual object over a virtual target. Based on Microsoft's design guidelines for mixed reality [Park et al., 2021], and prior research [Livingston et al., 2009, Erickson et al., 2021], the background object (e.g. virtual sign boards) or the target should be darker than the foreground objects (e.g. virtual text). A study by Erickson et al. [2021] shows that the user is more comfortable reading a darker text over a bright background.

Consequentially, we suggest that when designing a colourmap for OST-HWDs, all of its colour values should be brighter than the background–be it physical or virtual. Otherwise, if we apply a poorly designed colourmap to a virtual object, we may encounter the following issues: (1) difficulty mounting the virtual object over a brighter physical target, or (2) the virtual object becomes less comprehensible against a bright virtual background.

C4: Shader Operation and Programming

In theory, we can use post-processing shader blending operations to provide an enhanced AR user experience (See W3C [2015] for possible types of blending). For instance, if we want the user to only see shades of red of the physical world with a VST-HWD, we can implement a post-processing fragment shader that subtracts $RGB : (0, 1, 1)$ from the ambient light. In turn, this sets green and blue channels to 0 or less–effectively removing these colours from the environment. While this operation is simple to implement with VST-HWDs, this operation is more difficult with OST-HWDs. Although the user wearing an OST-HWD can directly see the physical world, the information on the ambient light of the physical world may be not fed into the shader. Therefore, the default background colour available to the post-processing shader is $RGBA : (0, 0, 0, 0)$. This limits the blending operation that we can perform. While an OST-HWD may be able to video-record the physical world and project the modified video back to the user, such a method requires complex transformation and may be too resource-intensive. Furthermore, using a video feed effectively transforms an OST-HWD into a VST-HWD.

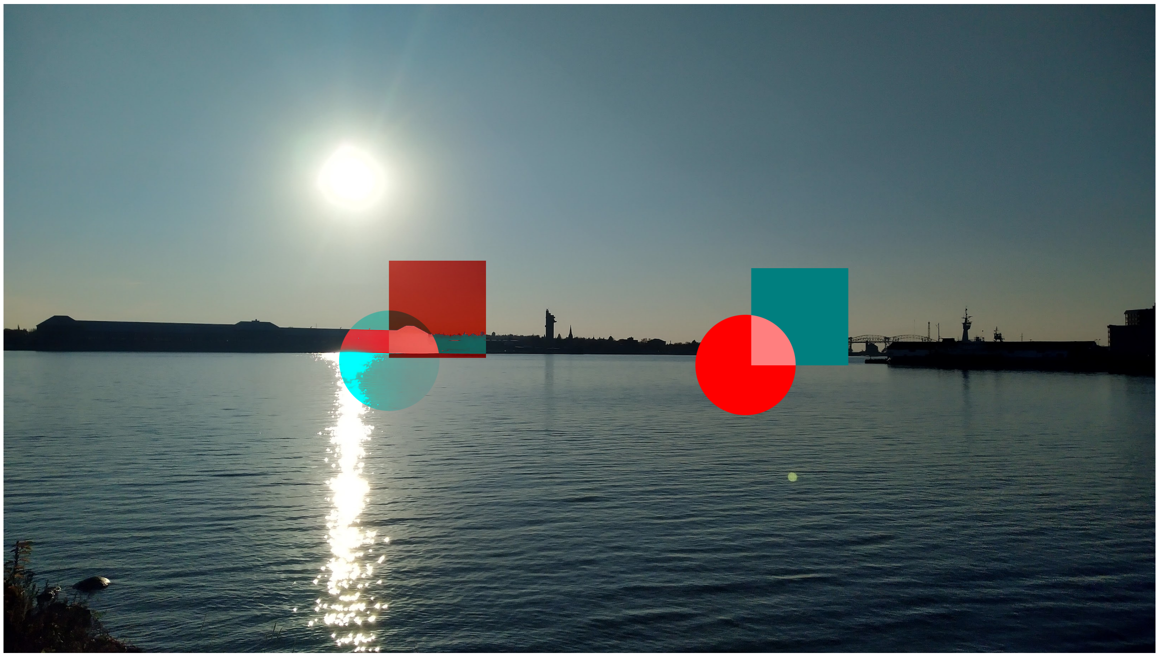


Figure 2.7: An inexperienced developer may expect the subtractive blending to resemble the set of a square and a circle on the left. However, the default subtractive rendering behaviour is on the right.

C5: Accessibility

Another factor that we must consider is accessibility. Some users may experience CVD which limits the colourmaps that we can use. For instance, those with green-red colourblindness cannot tell the green and the red colours apart. This means we are not able to use this scale.

2.5 Mobile Augmented Reality

Siriwardhana et al. [2021] define a piece of AR software as software that: (1) combines a set of real and virtual objects together in a physical world, (2) runs in real-time, and (3) tries to align virtual content relative to the real-world content. Mobile augmented reality (MAR) involves the use of a mobile device such as a cellphone or an untethered OST-HWD to show the virtual content. Since the HoloLens is an untethered OST-HWD and does not anchor the user in one spot [Kress and Cummings, 2017], we can consider it a MAR device. Meta 2 (Fig. 2.8)[2] is not a MAR device, despite its similar appearance to Microsoft HoloLens, because the device is tethered to a computer and the user must stay close to the machine [Pulli, 2017].

An OST-HWD is not perfect. It does have several issues. According to Itoh et al. [2021], while the device allows the user to have access to the physical world, the darkened visors can slightly distort the appearance of the physical world to the user. The headset also does colour inaccuracies which negatively affect the rendering of virtual objects [Itoh et al., 2021]. Furthermore, the field-of-view (FOV) for the virtual content is narrower (e.g. 52^o for HoloLens v2 [Williams and Ortega, 2021]) than the normal human FOV Simpson [2017]. This means if the virtual object is large and its angular distance is too far from the centre of the screen, it will appear truncated to the user.

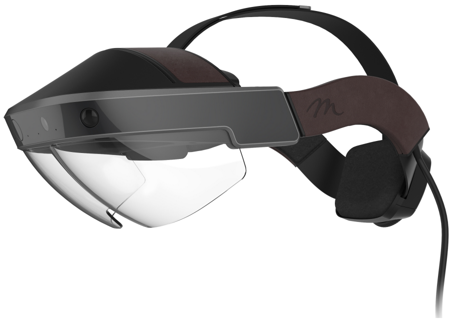


Figure 2.8: Meta 2, an OST-HWD similar to Microsoft HoloLens. However, unlike the HoloLens, the device is not a mobile AR device because it is tethered with a cable as seen in the figure. The image is from MetaMarket [2016].

In addition to OST-HWDs, there are other devices that we can use with MAR applications. One of the alternatives is to display the virtual content through a mobile phone. With this method, the mobile phone's camera streams a video live feed of the physical world. When relaying the streaming data to the user through the phone, the phone superimposes the virtual content onto the video stream. Since the phone relies on a video feed, the phone acts as a video see-through (VST) device. Unlike OST-HWD applications, phone-based MAR applications are readily available

[2] Not to be confused with Meta Company or the Meta Quest 2

to the public. Anyone can easily download and install on their phone. An example of this is Pokémon Go [Niantic, 2022]. More academic examples include AiR by Mathews et al. [2021]. AiR is a situated visualization application that visualizes pollution information by superimposing graphics on the streaming data. The application makes certain invisible air pollutants (e.g. carbon monoxide) more apparent to them. Mathews et al. [2021] did not evaluate the application.

Mobile phone-based MAR has less immersion and presence than other MAR systems. Immersion and presence, according to Liberatore and Wagner [2021], do not have single definitions. Rather, researchers tend to have their own definitions. In Liberatore's case, immersion is the level of sensory fidelity produced by the device. For presence, it is the subjective feeling of being in the virtual environment. Since phone-based MAR applications require the user to hold their phone and access the virtual content through it, the sense of immersion and presence is broken. OST-HWDs, on the other hand, allow the user to readily view the virtual content which augments the sense of immersion.

Another alternative to the mobile phone is using a VST-HWD. A VST-HWD is essentially a VR HWD that is capable of video-streaming the physical world and modifying the stream to include virtual content. A major disadvantage of this system though is the latency; there is a delay from video-recording the physical world and displaying it to the user. The latency can cause some users to feel cybersickness (e.g. nausea, headache) Freiwald et al. [2018]. Gruen et al. [2020] measured the latency rates of various VST-HWD devices (Modified Acer Mixed Reality, Oculus Quest 2, Oculus Rift S, and Valve Index). They found that humans have a mean baseline latency of 335 milliseconds. Meanwhile, the average latencies of the devices are from 394-434 milliseconds. Another disadvantage is that because a video recorder's FOV is narrower than the normal human FOV, the headset can limit the user's view of the physical world. A VST-HWD also blocks the user's face which can affect collaboration with other stakeholders. Lastly, a study by Ballestin et al. [2018] found that people tended to have better depth perception. Since VST-HWDs have multiple disadvantages that affect the perception of the environment, we decided to adopt Microsoft HoloLens v2, a type of OST-HWD, for our work.

2.6 Hybrid User Interface

A *hybrid user interface* involves using multiple types of interfaces together [Feiner and Shamash, 1991]. Gander is a type of hybrid user interface because it has an AR-based interface and a tablet-based interface that are being used together. It is important to note though that a hybrid user interface does not require the use of multiple pieces of hardware. For instance, Coninx et al. [1997] implemented a 2D/3D hybrid interface. While their interface makes use of two pieces of hardware (the pinch glove, and an immersive display), they argue that their interface is hybrid because the user relies on 2D dialogs to manipulate the 3D environment.

2.6.1 Input Device

Feiner and Shamash [1991] argue that AR should be used in conjunction with a separate input device since AR devices do not allow for precise input. Despite advancements in AR input methods (e.g. gaze tracking [Kress and Cummings, 2017]), imprecision is still an issue. Soares et al. [2021] conducted a study that compared handheld controllers, and gesture-based input methods found in HoloLens v2. They found that the participants were less precise without the controllers. Furthermore, the AR headset's hand gestures could be up to 2.5 cm off the targets. As such, having an extra input device is still beneficial. A tablet with high-precision support for touch gestures can serve as a companion input device for AR.

Multiple works [Surale et al., 2019, Drey et al., 2020, Beiner et al., 2022] explored the use of a tablet within the mixed reality environment to overcome the difficulties of manipulating virtual objects. Surale et al. [2019] developed TabletInVR, a tablet-based input system, to help with manipulating 3D virtual objects in VR. In their evaluation, they found the participants were able to understand the interface. Unlike Surale et al. [2019] who focused on the manipulation of 3D objects, the goal of VRSketchIn–implemented by Drey et al. [2020], is to support sketching within the VR. The stylus plus the tablet allow the user to draw in 3D. TabletInVR and VRSketchIn were evaluated with small sets of participants (n=6 per prototype); therefore, we cannot state the effectiveness of the techniques. PoVRPoint by Beiner et al. [2022] is focused on authoring presentations in VR. They evaluated the prototype by asking the participants (n=18) to perform tasks that could be found in presentation software (e.g. Microsoft PowerPoint). They found that overall, the participants enjoyed the experience. Unfortunately, effectiveness of the prototype was not measured. It is important to note that these prototypes were conducted with VR, instead of AR. However, the works demonstrate the effort of developing hybrid user interfaces with 3D virtual content and tablets.

There are multiple examples of hybrid user interfaces with AR in geospatial analysis. Tangible Globes by Satriadi et al. [2022] combines AR with a spherical device. Satriadi et al. [2022] designed the device to be used in three ways: (1) as a display device, (2) as a controller of a larger 3D virtual globe shown in AR, and (3) as a controller for a large AR-based 2D display. The prototypes were evaluated with four experts in geographic data visualization. Another is FieldView by Whitlock et al. [2020], which we have already discussed before. The user of FieldView uses a cellphone to perform immersive analytics tasks with AR in the wild, and in situ. Similarly to Gander, the cellphone has menus that execute commands for creating immersive 3D visualizations in AR.

2.6.2 Extending Display Device

AR can provide additional virtual displays to aid the users. Earlier works include providing small AR-based 2D widgets [Feiner et al., 1993] with 3D widgets [Di Verdi et al., 2003] that enhance the physical world, and contextualize physical objects. Pavanatto et al. [2021] argue that for using OST-HWDs to provide extra screens to a computer device. They state that as MAR devices, they do not take up space like physical screens. Despite OST-HWDs having display issues [Itoh et al., 2021], Pavanatto et al. [2021] found that the virtual screens are still useful for tasks involving desktop applications. STREAM by Hubenschmid et al. [2021] goes further by allowing the user to create room-sized 3D visualization. The user of STREAM can use the tablet to alter the visualization being displayed by an OST-HWD. Since the tablet is *spatially-aware*, it can track itself within the physical world and can aid the user with navigating the 3D visualization. MARVIS by Langner et al. [2021] is an example of using AR to enhance the tablet. However, instead of creating a room-size visualization. The AR is used exclusively to support 3D visualization above the tablet.

STREAM and MARVIS have not been extensively evaluated. Therefore, the benefits of the systems are unclear.

OST-HWDs can provide extra virtual screens, and add support for 3D visualization to 2D display devices. However, they are not without issues. OST-HWDs have colour display issues that we must be cognizant of. Furthermore, OST-HWDs may have lower display resolutions than the tablets. For instance, Microsoft HoloLens v2 has the default display resolution of 1440x936 px. [A. Turner, v-chmccl, and V. Tieto, 2022] while Microsoft Surface Book 3 (15in Screen) has the resolution of 3240x2160 px. [Microsoft, 2022]. Using two levels of display resolutions means an AR+tablet interface may become a focus+context (F+C) interface. Baudisch et al. [2001] describe F+C as using two types of resolutions: the focus resolution, and the context resolution. The focus resolution is a higher one, and it supports the user looking at content in detail. Meanwhile, the contextual resolution is lower, and it is for the user to glean contextual information. The focus area is smaller than the contextual one. It is important to note that we do not need multiple display devices to implement F+C. For instance, the original implementation of F+C by Baudisch et al. [2001] simply uses a single large monitor with the focus area being rendered better than the context area. Furthermore, having two resolutions is not sufficient. For an AR+tablet to be a true F+C interface–with the tablet representing the focus resolution and the AR representing the contextual information, the user must actively be focusing on the tablet and only use the AR for obtaining contextual information. Therefore, while Gander's bi-resolution displays support F+C, it is not necessarily enforced. If the user decides to mainly rely on the AR and mainly uses the tablet for input, it does not conform to the F+C paradigm.

2.6.3 Augmented Reality Content Placement

There are multiple ways of aligning AR content with other hardware. For example, Satriadi et al. [2022]'s Tangible Globes, which we have discussed before, allow AR content to be placed onto the spherical input device or onto the physical world itself. Tönnis et al. [2013] call this concept "mounting." For Gander, the AR content is mounted on top of the tablet by default. However, during the synoptic study, we observed some participants separating the AR content from the tablet by panning the map so much that the AR map became detached from the tablet. While we did not anticipate this behaviour, the outcomes show that Gander could also support mounting to the physical world.

Another concern about the placement is navigation. If we mount the AR content onto a physical object, the AR content should also move along with the physical object. An example of this is MARVIS by Langner et al. [2021] which uses OST-HWDs to add 3D AR content on top of tablets. When the user moves the tablet, the AR content moves along with it. While HoloLens v2 supports tracking of objects in the physical object through its cameras, we found that the latency of the cameras is too great. Therefore, we did not implement object-following. Instead, we disallowed the user from moving the tablet–effectively, turning the tablet into an immobile kiosk once the mounting process is complete.

2.7 Concluding Remarks

To develop and evaluate Gander, we researched prior literature in many areas. First, we studied the steps, activities, and concepts involved in MLR. We applied the first principles that we distilled from the literature to the workflow of Gander itself. Then,

Topic	Influenced
Machine Learning and Interactive Machine Learning	Research in the topic helped us to determine the main tasks for the vertical slice, and the type of statistics used.
Exploratory Data Analysis	The concept of data exploration inspired the creation of the pre-fit and the post-fit stages.
Information Visualization	The colourmaps were created based on information visualization guidelines. The work on this topic also helps with glyph designs and the determination of glyphs used in the synoptic study.
Immersive Analytics	The work influenced the design of Stacked, and room-sized displays. Furthermore, it informs colourmap designs for Microsoft HoloLens v2. The future designs of Gander (Ch 9) plans to further incorporate immersive analytics.
Hybrid User Interface	The work helped us to consider how multiple devices could be used in conjunction.

Table 2.1: The summaries of how certain fields that we reviewed influenced the development of Gander, and our research.

we analyzed the literature on immersive analytics. As a transdisciplinary field, we also researched topics relevant to immersive analytics such as information visualization, and visual analytics. Furthermore, we explored works in geospatial analysis. This led us to adopt the glyph-based visualization as the main visualization system for Gander. Lastly, we analyzed how multiple mobile devices could be combined to complement each other's strengths and weaknesses. For instance, OST-HWDs support large-area displays, but have limited input while a tablet, a device with a smaller screen, supports a superior touch-based input experience. Table 2.1 summarizes how the fields influence ourr research.

2.7.1 Research Gaps

Based on these works, we identified the following research gaps.

Immersive Analytics for Interactive Machine Learning

Even though Gander is not a full IML system like Orange, we argue that Gander can serve as a stepping stone toward a full IML system with a hybrid AR+tablet interface. At this point, research in IML mainly focuses on desktop-based software. Whitlock et al. [2020] point out that there is a need for an immersive analytics system that can work in-situ and can provide a large room-sized visualization. Therefore, we argue for Obj1, or the aim to develop a vertical slice prototype that can somewhat represent an ideal IML system in the future. While our work is limited to the domain of geospatial analysis, future researchers can extend Gander to suit their own uses through data spatialization.

Lack of Empirical Studies for Immersive Glyph Visualization

Although there is a large body of work on glyph comprehension within the context of information visualization and visual analytics (e.g, Jankun-Kelly et al. [2010], Peña-Araya et al. [2020]), AR glyphs are typically evaluated holistically as a part of the system. Therefore, the evaluation is highly qualitative and difficult to generalize. Without a good empirical understanding of AR glyphs and systems, we cannot fulfill Obj2–i.e. understand how the user makes use of a glyph in an immersive IML system.

Lack of Empirical Studies for AR+Tablets Hybrid User Interfaces

The concept of combining a mixed reality interface with a tablet one is not new. However, like AR-based glyph visualization works, most of the literature in this area focuses on providing system descriptions rather than on evaluating the effectiveness of the interface. While some papers describe experiments (e.g. Beiner et al. [2022]), most rely on qualitative studies with small sample sizes (e.g. $n = 4$ in Satriadi et al. [2022]) – or not having any evaluation at all (e.g. Lau et al. [2019]). To ensure that our work minimizes these gaps, we develop human-participant studies that provide some empirical evidence on the effectiveness of an AR+tablet interface. This work fulfills Obj3–i.e. designing and evaluating an AR+tablet hybrid user interface.

Chapter 3

Exploratory Background Work

As immersive analytics is a broad and interdisciplinary domain which involves research from multiple areas [Ens et al., 2021], we must first identify the area for research. In our exploratory work, we performed research in the following areas: (1) visual query language design, (2) out-of-view target acquisition, and (3) variance structure visualization. When working on these areas, we found that they required a functioning system. Hence, we shifted our focus towards Gander. In the end, designing and evaluating Gander became a more important enterprise. As such, we shifted our focus from these areas. Still, elements of these works were incorporated into the design and evaluation processes of Gander.

In addition to the exploratory research, we also worked with the Government of Nova Scotia to analyze online survey data. The analysis was largely qualitative in nature; however, it contained elements of spatial analysis. Since the survey was about amalgamating two jurisdictions into a single one, our work involved identifying the differences between the two. This work also involved the use of EDA and data visualization. Our experience from this project inspired the basic design of Gander.

3.1 Explatory Work within Computer Science

3.1.1 Visual Query Language

"Mimi"[1] was an early visual query language for an immersive analytics system. The system allowed the user to combine multiple visual widgets to form a query; a low-fidelity prototype of this is available in Fig. 3.1. In this example, the user is inquiring if a ship is too close to an ice floe or not.

We applied semiotics when designing Mimi. We considered how the arrangement or the syntax of the visual widgets can influence the user understanding. Furthermore,

[1] Mimi is based on "Mimir", the name of the project. It is also based on the Japanese word for "ear."

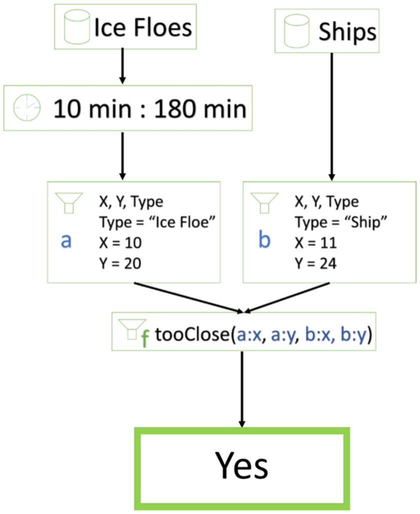


Figure 3.1: A prototype of the visual query language showing a query which checks if a ship is too close to an ice floe.

we consider the graphical representation of the glyphs used for Mimi. Mimi icons are representations that partially resemble the original object [Chandler, 2018]. For instance, the boat icon in Fig. 9.8 resembles a real ship that it aims to represent. The use of semiotics in human-computer interaction (HCI) is usually constrained to specific areas of research. For instance, we tend to focus on designing visualizations [Borgo et al., 2013] or a user interface [Barr et al., 2004]. However, semiotics could also be applied to different senses. For instance, we could apply semiotics principles to analyze language translation.

Related to Mimi, is the "average-based selection." An average-based selection is a type of query that is less precise than a normal database query. Instead of selecting data that meet specific criteria, we choose data based on the group average and ignore data whose attributes are too extreme. In order to design an average-based selection technique, we started to design an immersive analytics system (Fig. 3.2). Since an average is a type of statistic, we envisioned the system must also be able to perform some type of statistical inference. For instance, if we have means from multiple selections, then we should be able to perform t-tests on them. From this idea, we evolved the prototype to become Gander, a system that possesses elements of IML, and is capable of statistical inference (namely, MLR).

During the design and the development of Gander, we realized that implementing Mimi and the "average-based selection" requires better understanding of how the user

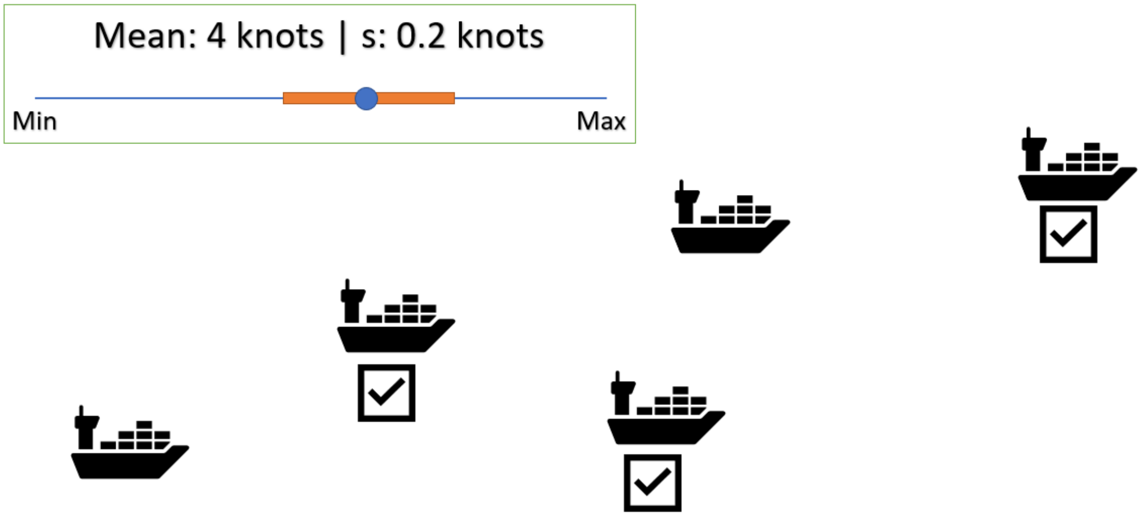


Figure 3.2: A prototype for average-based selection. On average, the ships are travelling at 4 knots. The ships that are travelling between 4 ± 0.2 knots are selected and have checkmarks annotated to them.

perceives the visualization. Therefore, we excluded them from the design. Elements of this work later return in the chapter describing proposed modification for Gander (Ch. 9).

3.1.2 Out-of-view Target Acquisition

Another topic that we explored is out-of-view target acquisition. This topic is important because room-sized interfaces can have many out-of-view targets. This can affect how the user operates within Gander. Therefore, we conducted two studies prior to designing Gander. These studies involve comparing multiple visual cueing techniques for out-of-view targets. We designate these studies as the visual cueing study 1, and the visual cueing study 2. In the visual cueing study 1 (described in [Hu et al., 2021]), we compared our two techniques: fSOUS (Fig. 3.3-a), and bSOUS (Fig. 3.3-b). FlyingARrow (Fig. 3.4-b) by Gruenefeld et al. [2018] was used as the control technique. fSOUS (faint Sign Of the UnSeen) is subtle. The cue is faint black graident. Meanwhile, bSOUS (bold Sign Of the UnSeen) is very explicit, appearing as a red blinking ring. We thought fSOUS could be effective despite its subtle design. As Bartram et al. [2001] state, animation should still be noticeable within the peripheral vision; however, the study demonstrates that bSOUS and FlyingARrow are generally more effective in an out-of-view target selection task using head gaze (Fig.

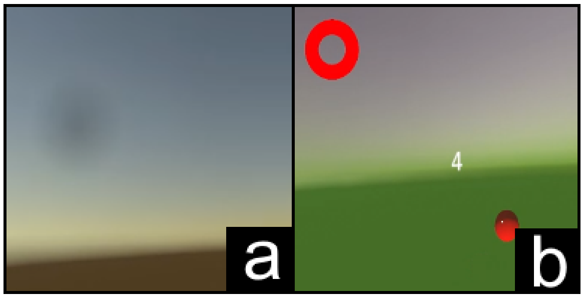


Figure 3.3: The two visual cueing techniques. **a:** fSOUS. **b:** bSOUS.

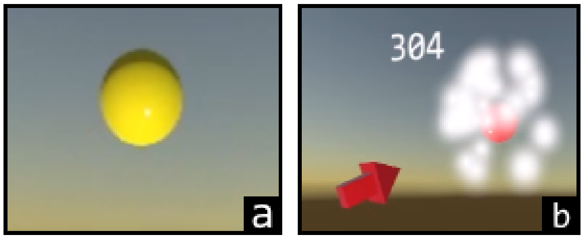


Figure 3.4: The screenshots of the game used in the study. **a:** Head-gaze is used to start the target acquisition process. **b:** To complete the acquisition of the target, the participant must dwell on the target–after which a sparkling effect is played. The arrow represents the FlyingARrow technique.

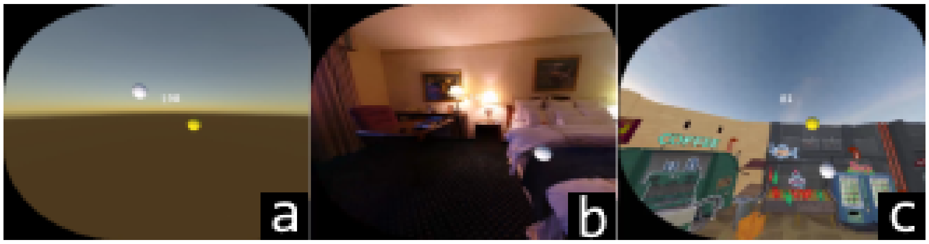


Figure 3.5: The environment used in the study. **a:** The control environment, which is uncluttered. **b:** The realistic environment, which is moderately cluttered. We used resources from Persson [2013] **c:** The 3D environment, which is the most cluttered. We used resources from Pulpil Labs [2021], Persson [2013]

.

3.4). Furthermore, our study varies the virtual environments (Fig. 3.5) to determine whether visual clutter could have any negative impact. We found that fSOUS was more affected due to its less noticeable design. The control technique, FlyingARrow, turns out to be slower than fSOUS and bSOUS; the speed of the arrow itself influenced the participants' speed.

In the visual cueing study 2 (also described in [Hu et al., 2021]), the participants used modified versions of FlyingARrow (Fig. 3.6), a visual cueing technique developed by Gruenefeld et al. [2017]. The technique used a 3D arrow that flies from the

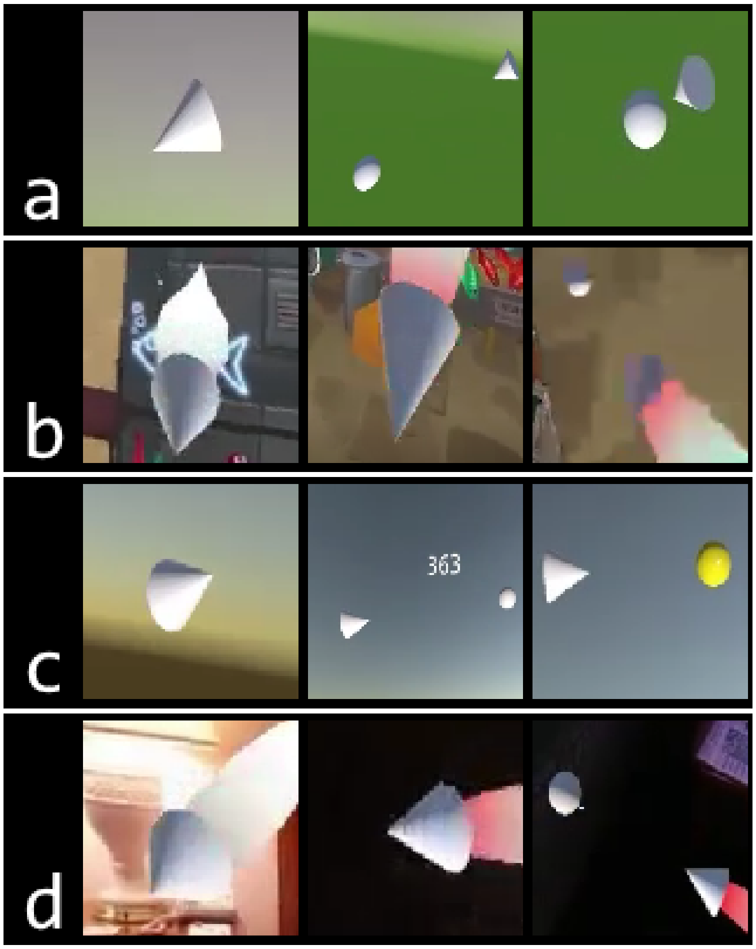


Figure 3.6: The variations of FlyingARrow (FA) **a:** FA-Arc-Trail (Control, the arrow flies directly to the target without any trail) **b:** FA-Arc+Trail (The arrow files directly to the target while emitting a trial) **c:** FA+Arc-Trail (The arrow files in an arc around the user to the target). **d:** FA+Arc+Trail (The arrow files in an arc to the target with a trail).

.

origin toward the out-of-view target. Unlike the techniques in the first study, these techniques did not persist in a participant's peripheral vision. Therefore, there was a possibility of the participants being unable to follow the cue. The modifications included: the use of visual trails, and making the arrow fly in a curve around the user instead of a straight line. The trails make FlyingARrow more visible to the user, similar to fSOUS and bSOUS which are always visible to the user. Meanwhile, the orbiting curve makes the arrow's position relative to the user than to the physical world which makes the techniques more similar to fSOUS and bSOUS. We found that the trial was helpful in improving speed, but the curved trajectory was not. Both techniques require additional adjustments to make them more comfortable to use.

In the end, Gander did not incorporate any technique to alert the user to out-of-view targets, because we must first identify the tasks that require acquiring out-of-view targets. This suggests that a vertical slice must first be created.

3.1.3 Variance Structure and Uncertainty Visualization

Prior to adopting likelihood as the main measure of goodness-of-fit, we considered using variances. Originally, Gander glyphs would display individual variance instead of individual likelihood in the post-fit stage. To understand better understand variances, we researched the work on variances. Example topics include: variance estimation (e.g., Zientek and Yetkiner [2010]), variance-based effect size R^2 (e.g., Cohen [1988], Lewis-Beck and Skalaban [1990], Rights and Sterba [2020]), variance visualization (e.g., Parker et al. [2014]). We also researched how variances could be "split" at the level of an individual datum. When we were trying to establish the mathematical soundness of "splitting" a variance, we came to a realization that likelihood is a more flexible approach for expressing goodness-of-fit.

Although we no longer use the variance-based approach, the research in variance is still relevant for trajectory analysis. As it turns out, mean-squared distance (MSD), a type of measurement for trajectory, is a type of variance itself. Therefore, we apply knowledge in variance to the analyses in the synoptic study (Chapter 5) for the analysis of mean-squared distances (MSD). We develop Appendix A which expands upon the MSD analysis method used in the synoptic study and provides additional technical backgrounds.

3.2 Transition to Immersive Analytics

The exploratory work, except for the one on out-of-view target acquisitions, does not require immersive analytics to progress. For instance, "Mimi" can also work with a standard desktop computer with a mouse. However, immersive analytics allow us to better explore in more creative directions. "Mimi" widgets, instead of being clicked on and dragged around by mouse, could instead be grasped and moulded by the user's hands. The user could also explore variance structures either through the bird's eye view or as a virtual installation that the user could walk in between.

Thinking how the exploratory work can fit within a single immersive analytics is difficult, and requires a foundational immersive analytics system. As such, we made the decision to first develop Gander as a vertical slice prototype. With a vertical slice, then we can use the work to expand the slice. We found the visual query language can be helpful for improving the interactivity of Gander–particularly for data selection, and performing statistical tasks. The work on variance structure is less helpful, because we found likelihoods are more flexible. Meanwhile, out-of-view targets may supplement the visual query language as a way to help the user to identify out-of-view widgets.

3.3 Miscellaneous Work with Windsor and West Hants, Nova Scotia

"Windsor/West Hants Together" was the name given to the amalgamation process that merged Windsor and West Hants, Nova Scotia, Canada into a single municipality [The Government of Nova Scotia, 2020]. During the process, the residents of Windsor and West Hants completed an online survey. We analyzed the survey data, and made multiple recommendations that formed the basis of the amalgamated municipality based on the data. The most notable one is to simply name the new municipality "West Hants", and not with other combined names such as Windsor/West Hants. This recommendation was made due to most of the respondents wished for a simpler name. Another major result that we observed was that both Windsor and West Hants residents wanted more cooperation, and more efficient management of resources.

The analysis was largely qualitative in nature. However, since the data came from three populations: (1) the residents of West Hants, (2) the residents of Windsor, and (3) outside residents, we argue the analysis contained elements of geospatial analysis. While we ultimately settled for qualitative analysis, we also performed an EDA using map-based data, and spent time finding applicable quantitative techniques. Furthermore, we also tried to detect any difference between the residents of the populations–even though no tangible difference was found in the end. The only major difference was that more Windsorites wished to have "Windsor" as the name of the new municipality.

As a part of the EDA, we used map-based visualization to initially identify where the respondents were from. The survey question allowed the respondents to provide their addresses. Based on the EDA, we were able to deduce that there were three populations. We were also able to learn certain characteristics of the areas. For example, West Hants is more sparsely populated than Windsor. The EDA also allows us to understand the issue named "the Crossing." We found that certain facilities (e.g. a school, and a hospital) close to the border of Windsor and West Hants were poorly connected. The West Hants residents living close to the landmarks must take a long detour in order to access them by road.

The EDA work that we performed here helped to partially form the basis of Gander. We learned the importance of visually inspecting geospatial data using map-based visualization. Furthermore, looking at the map and the individual addresses in detail allowed us to gain some insight. Even if we reported aggregated data in the end, knowing individual data allows us to select the best methods of aggregation and analyses.

Chapter 4

Design of Gander

Gander is an AR+tablet interface for exploratory multiple linear regression analysis. It relies on glyph-based and map-based visualization to present information. The design intent is to support data exploration on both the original data, and the likelihoods of the models fitted from the data. Using room-sized visualizations, the user can benefit from being able to navigate through the glyph fields themselves. The glyph-based visualization allows the user to see individual data points and likelihoods. Gander has three main stages: data selection, pre-fit and post-fit. In the data selection stage, the user selects a dataset from a given list that they wish to use in their study. Pre-fit is for analyzing data and for model selection. The user can view a map and a glyph field, and then compare the variables. Post-fit is similar to pre-fit; however, the likelihood information of the models is displayed instead.

When we develop Gander, we follow the *Design as Research* approach. According to Stapleton [2005], design as research is a process where we use design to better understand the overall research topic, before researching it. Design as research is necessary when the topic itself is nebulous. For instance, when Stapleton [2005] was researching designing an educational game, he found that he could not immediately apply traditional research methods, like ethnography or interview. Instead, he had to start to design the game with his collaborators. Once a prototype had been created, research could begin. Herriott [2019] outlines design as research as having the following steps: (1) background research, (2) creation of an object based on the research, (3) analysis of the object, and (4) creating new theories. For our work, we first utilize design as research to create Gander, and then use Gander to conduct experiments and studies. Despite Fisher's push towards mechanizing statistical practices into concrete steps [Ziliak, 2008], statistical practices can still significantly vary. While MLR is deemed as a simple statistical method, it is still replete with different sets of recommendations.

Gander is a vertical slice prototype. Unlike a conventional prototype in HCI, a vertical slice is a usable one. However, it does not have all functionalities [Ratner and Harvey, 2011]. For instance, Gander can open data files, visualize the data, fit a model, and display the model's likelihood; however, all the steps have limited functionalities. An example of the limitations is that Gander can only perform MLR– even though there are many other types of regression. The goal of a vertical slice is to present a vision of the complete system without being distracted by having to implement and test various other functionalities. Further, several of Gander's functionalities do lend themselves well as low or medium-fidelity prototypes. For instance, Stacked, one of the glyph-based visualization techniques, relies on the parallax effect which can be difficult to explain using paper-based or 2D flat prototypes.

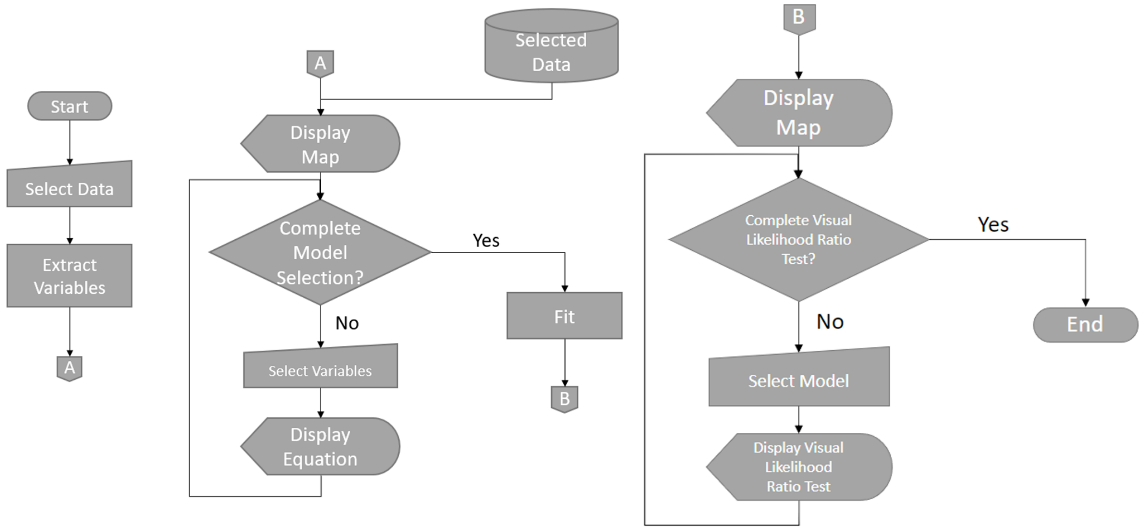


Figure 4.1: Flowchart describes how the user interacts with Gander. The first column represents the data selection stage. The second column represents the pre-fit stage where the user select variables to visualize. After the visualization, the user can keep changing the variables until they are satisfied. The third column represtns the post-fit stage where the user examine the likelihood of a model against another one.

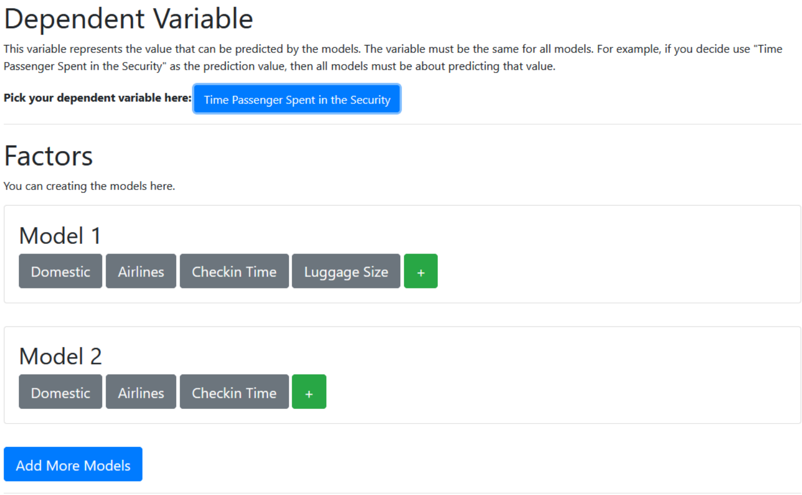
Gander possesses a workflow which has three main parts: (1) data selection, (2) pre-fit, and (3) post-fit. The data selection stage involves the user choosing the data that they want to work with. In the pre-fit stage, the user performs EDA and feature selection. In the post-fit stage, the user explores likelihood information. Fig. 4.1 summarizes the general workflow of Gander. Section 4.2 provides additional information.

The design rationale of Gander is to assist the user in seeing details at the most granular level. This allows the user to see patterns that they may otherwise miss. For instance, if we use a cartogram representing averages of an area, the cartogram can hide specific patterns that may be worthy of additional exploration. This approach is different from the usual approach where the user always sees the aggregated information, such as averages and effect sizes for the whole model (e.g. R^2, likelihood ratios). This approach is to avoid ecological fallacy–i.e. generalizing aggregated information to individual data points [Salkeld and Antolin, 2020]. An example of ecological fallacy is when we insist that a voter of a particular party lacks a college degree, because voters of that party tend to not have post-secondary education.

4.1 Past Designs

Gander has gone through multiple rounds of evolutions. Although we only evaluated the high-fidelity version of Gander, we developed several low- and medium-fidelity prototypes which we describe here.

4.1.1 Airseer



Dependent Variable

This variable represents the value that can be predicted by the models. The variable must be the same for all models. For example, if you decide use "Time Passenger Spent in the Security" as the prediction value, then all models must be about predicting that value.

Pick your dependent variable here: Time Passenger Spent in the Security

Factors

You can creating the models here.

Model 1

Domestic | Airlines | Checkin Time | Luggage Size | +

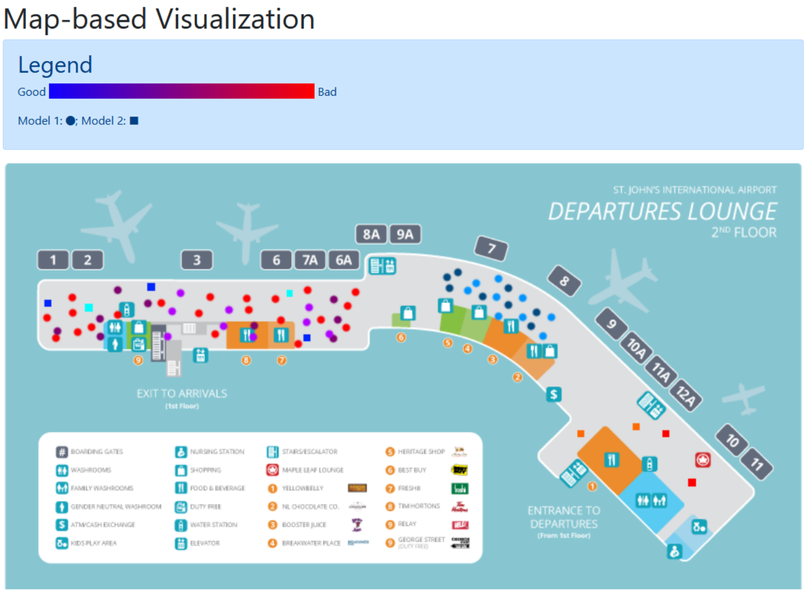
Model 2

Domestic | Airlines | Checkin Time | +

Add More Models

Figure 4.2: Airseer's pre-fit prototype.

Initially, Gander started out as an unnamed low-fidelity prototype only meant for data selection using immersive technologies. This prototype was described in Chapter 3. As we became interested in incorporating IML features into the prototype, we created 'Airseer' during an aviation-themed Hackathon. Unlike the unnamed prototype, Airseer's primary focus was solely on statistical analysis. Therefore, it



Map-based Visualization

Legend

Good Bad

Model 1: ●; Model 2: ■

Modified from: https://stjohnsairport.com/wp-content/uploads/2019/09/YYT-Map-interior-SecondFloor-Sept2019-e1568812696234.png

Figure 4.3: Airseer's post-fit prototype. The prototype includes a screenshot from [St. John's International Airport, 2019].

was developed as a desktop web-based application rather than a MR one, and it did not include any exploratory feature. The user started out in the pre-fit stage (Fig. 4.2). However, the user could not perform any data exploration at this stage. They also could not pan the map. Instead, the user could only visualize the information in the post-fit stage (Fig. 4.3).

Due to the Hackathon being focused on airport-based activities, Airseer was supposed to work with airport data. However, obtaining airport data was difficult. Therefore, we decided to turn Airseer into a general-purpose geospatial application.

4.1.2 Early Versions of Gander

After Airseer, we decided to add the use of immersive technologies, such as AR. This would make the project more compatible with our exploratory work with out-of-view target acquisitions Hu et al. [2021]. We then combined elements of the unnamed prototype and Airseer together to create Gander. We chose this name for two reasons. First, we named it after the expression "to take a gander." Secondly, Gander is also the name of a Canadian town famous for its airport. Therefore, this name serves to remind us of its original purpose as aviation software.

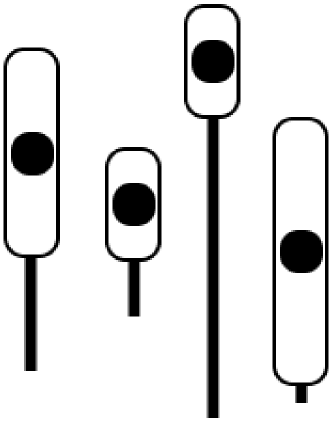


Figure 4.4: Prototypes of the punching bag plots

We envisioned the users to be explorers who were spelunking through the environment, in order to understand it. In essence, the early version of Gander would be a Digital Earth application. We proposed multiple types of data visualizations before settling down on the glyph-based visualization. The examples include using the lollipop plot where each lollipop represents a univariate variable, generating a raster texture similar to [Kumar et al., 2013], and generating relief-shear terrain maps based on the data and on the information. We decided to adopt the glyph-based visualization since it is more suitable for visualizing single data points. While the lollipops also visualize single data points, they can be more prone to overdrawing.

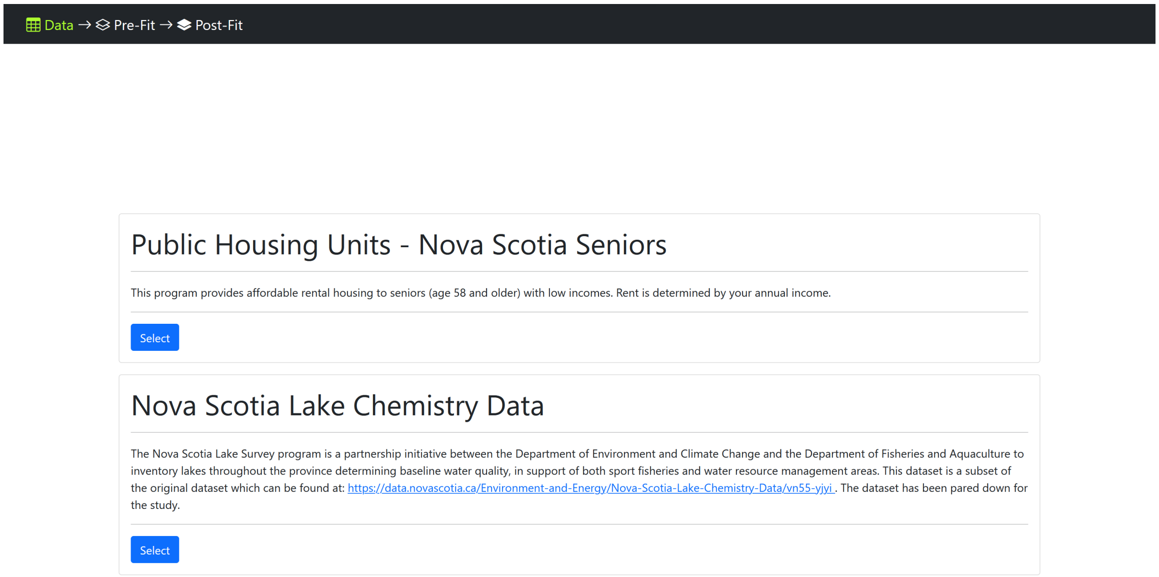
We also considered a "punching bag" plot (Fig. 4.4), a modification of the lollipop plot. Essentially, a punching bag is a lollipop modified to convey a confidence interval. It is similar to a point-and-whisker plot Cumming [2014]. However, since we decided not to proceed with the lollipop plot due to potential difficulties of representing multivariate information, and overdrawing.

The design for Gander started to solidify when we determined that we would focus on combining AR together with a tablet interface. We scaled down the level of immersion. Instead, we focused on using AR to extend a tablet's interface. Furthermore, we focused on implementing a vertical slice prototype. This meant focusing on the core features that would allow the user to start with data selection and end with an analysis after fitting the data set. Less emphasis was put on developing an immersive Digital Earth application. Originally, the post-fit stage was supposed to be based on R^2 or other variance-based effect sizes such as Cohen's f^2. However, we later found that likelihood is more flexible.

4.2 Workflow and Interaction Stages

In general, the steps could be broken into two: pre-fit stages and post-fit stages. Prior to reaching these stages, the user must first calibrate the device. Then, they must select the data. The data selection is performed through the tablet itself. This section focuses on the workflow of Gander without providing details of implementation. For implementation details, and more information on calibration, please refer to Section 4.3.

4.2.1 Interaction Stage 1: Data Selection



Public Housing Units - Nova Scotia Seniors

This program provides affordable rental housing to seniors (age 58 and older) with low incomes. Rent is determined by your annual income.

Nova Scotia Lake Chemistry Data

The Nova Scotia Lake Survey program is a partnership initiative between the Department of Environment and Climate Change and the Department of Fisheries and Aquaculture to inventory lakes throughout the province determining baseline water quality, in support of both sport fisheries and water resource management areas. This dataset is a subset of the original dataset which can be found at: https://data.novascotia.ca/Environment-and-Energy/Nova-Scotia-Lake-Chemistry-Data/vn55-yjyi . The dataset has been pared down for the study.

Figure 4.5: Data selection screen. The data sources are available on the screen.

After the calibration process, Gander presents a list of maps that the user can select. A map has two files associated with them: a table file, and a metafile. The table file contains spatial information (e.g. longitude and latitude) that Gander uses to place the glyphs, and information for the variables. The data's scaling type must be numerical (e.g. absolute or ratio) so that Gander can normalize the data to be between zero and one. The normalization process is important for assigning colour values to the glyphs that represent the data. A categorical variable or feature is acceptable, if encoded through procedures such as dummy encoding. A metafile is a description file that instructs Gander on how to render the map. It contains information on the map boundary, and the scaling parameters. Once the user selects a desired data source, Gander applies the metafile information onto both the tablet's map display, and the AR interface. For the full details, please refer to Section 4.3.1.

4.2.2 Interaction Stage 2: Pre-fit

At this stage, the user engages in data exploration so that they can better understand the data, and their distributions on the map. In general, using the tablet, the user can visualize variables as AR glyphs. The user can also pan the map around using touch gestures. We did not implement zooming, because it requires additional design considerations. For instance, Müller et al. [2014] state that when glyphs are zoomed out, they can display fewer details than when they are zoomed in.

Based on our first principle analysis of prior literature on MLR and [Friedrich, 1982, Braumoeller, 2004, Daoud, 2017, Zhang et al., 2009, Dudley and Kristensson, 2018, Jenkins-Smith et al., 2021], the user should perform the following:

- **Parsimony.** Does a model have too many independent variables (IVs)? If multiple IVs are highly correlated, some should be removed. Otherwise, a fitted model will not be parsimonious and can have issues like multicollinearity [Daoud, 2017].

- **Multiplicativity.** Considering if there is any multiplicavity (a.k.a interaction effect) between the IVs, or if each IV's effect is independent. Not considering multiplicativity can affect a model's accuracy [Braumoeller, 2004, Friedrich, 1982].

- **Correlation.** The IVs must be able to explain the variances of observed DV values. Example effect sizes include R^2 and adjusted R^2 to quantify this [Lewis-Beck and Skalaban, 1990]. In visualization, if there is a correlation between two or more variables, a change in one variable should also be observable in another.

For instance, when the values of one variable are observed to be increasing, the values of another variable could be increasing or decreasing. Correlation may not exist if the values of the first variables do not predict the values of the others.

- **Spatial Autocorrelation.** Data may cluster spatially, which may complicate fitting; for instance, overall national statistics can differ at state-level Peña-Araya et al. [2020].

Tablet Interface

After having selected a map, the tablet presents the user with Fig. 4.6. The user can pan the map using touch gestures. On the top, there is a navigation bar (Fig. 4.7. The left side of the bar indicates the current stage (pre-fit). On the right side, there are buttons that the user can tap on to launch the Variable Picker and the Equation Modeller. The user can pan the map itself using swipe gestures. Doing so will also move the AR content since it is registered to the tablet.

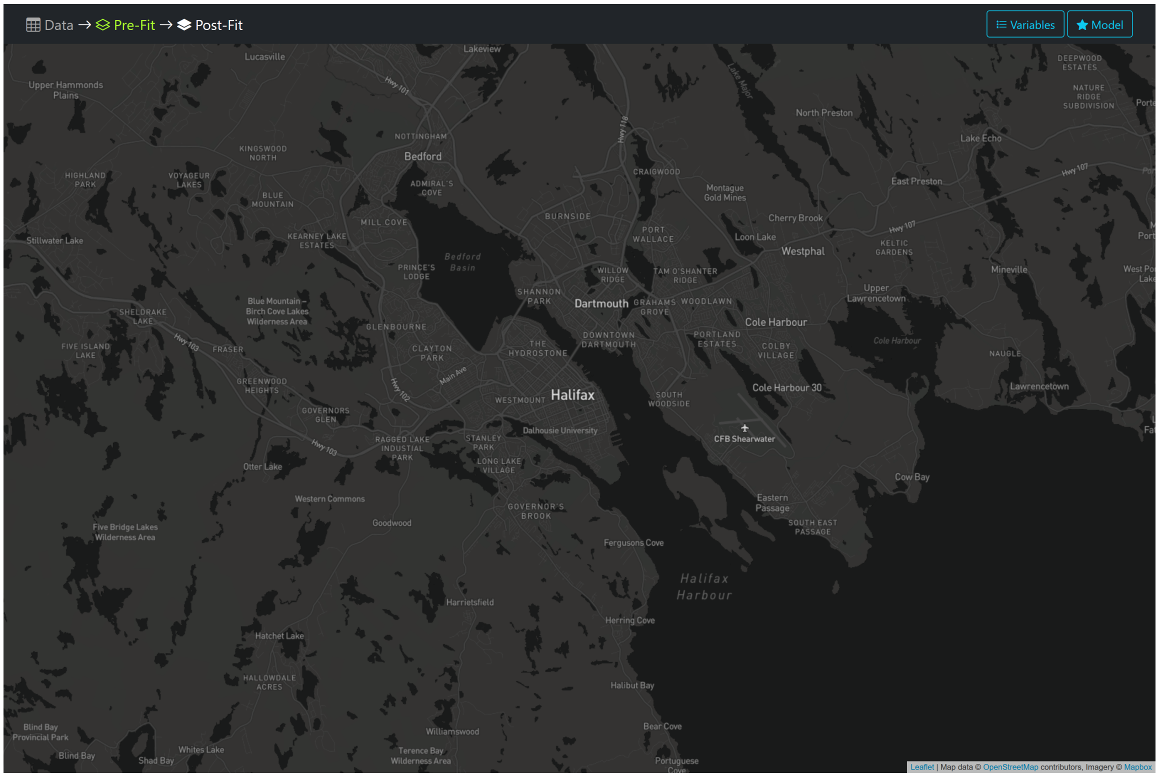


Figure 4.6: Gander in the pre-fit stage on the tablet at the beginning of the pre-fit stage.

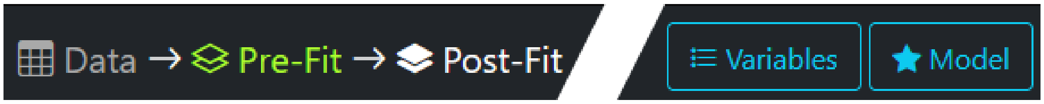
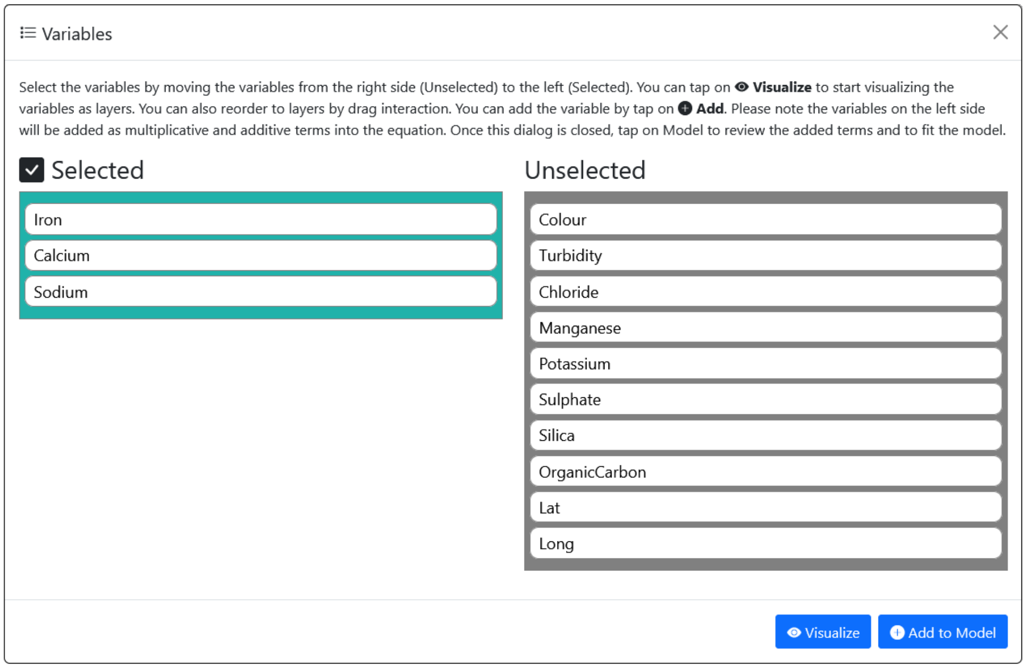


Figure 4.7: The navigation bar in the pre-fit stage. The centre empty section of the bar has been removed to improve presentation.


Variables

Select the variables by moving the variables from the right side (Unselected) to the left (Selected). You can tap on **Visualize** to start visualizing the variables as layers. You can also reorder to layers by drag interaction. You can add the variable by tap on **Add**. Please note the variables on the left side will be added as multiplicative and additive terms into the equation. Once this dialog is closed, tap on Model to review the added terms and to fit the model.

☑ Selected

Iron
Calcium
Sodium

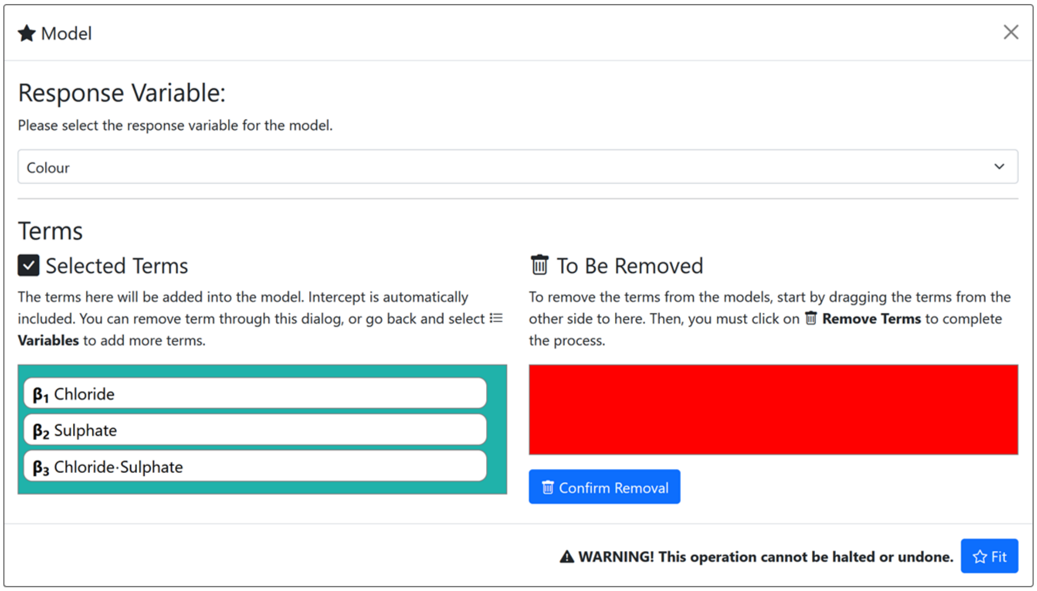
Unselected

Colour
Turbidity
Chloride
Manganese
Potassium
Sulphate
Silica
OrganicCarbon
Lat
Long

Figure 4.8: The Variable Picker. The variable names are properties found in lakes of Nova Scotia because we used the Nova Scotia lake chemistry data [The Government of Nova Scotia, 2021].

The Variable Picker dialog box (Fig. 4.8 supports two main operations: displaying the glyphs, and adding the variables into a model. To display glyphs, the user can drag and drop the variables from the “Unselected” column into the “Selected” column. Then, they can rearrange the variable through swipe gestures to order the Stacked glyphs in AR. Once the user taps on the “Visualize” button, the glyphs are displayed in AR. If the user wants to add the selected variables to the model, they can tap on “Add to Model.” The Variable Picker can be closed and re-opened as many times as the user would like. Please refer to Section 4.2.2 for more information on how the user performs data exploration with the glyphs. That particular section describes how the user can interpret glyph information. It is important to note that the tablet interface does not display any glyph.

After having selected the variables, the user can review them through the Equation



★ Model ×

Response Variable:

Please select the response variable for the model.

Colour

Terms

☑ Selected Terms

The terms here will be added into the model. Intercept is automatically included. You can remove term through this dialog, or go back and select ☰ **Variables** to add more terms.

β_1 Chloride

β_2 Sulphate

β_3 Chloride·Sulphate

🗑 To Be Removed

To remove the terms from the models, start by dragging the terms from the other side to here. Then, you must click on 🗑 **Remove Terms** to complete the process.

🗑 Confirm Removal

⚠ **WARNING! This operation cannot be halted or undone.** ☆ Fit

Figure 4.9: The Equation Modeller.

Modeller. On the navigation bar, the user taps on “Modeller” to open the dialog (Fig. 4.9). The dialog contains the list of variables selected as well as the multiplicative variables. E.g. in Fig. 4.8, *Iron*, *Calcium*, and *Sodium* are selected. Assuming the user decides to add them, in the Equation Modeller will show these variables as well as *Iron x Calcium*, *Calcium x Sodium*, *Iron x Sodium*, and *Iron x Calcium x Sodium*. The touch-based drag-and-drop allows the user to remove undesired variables. Once the user is ready, they can tap on “Fit” to create a MLR model.

Augmented Reality Interface

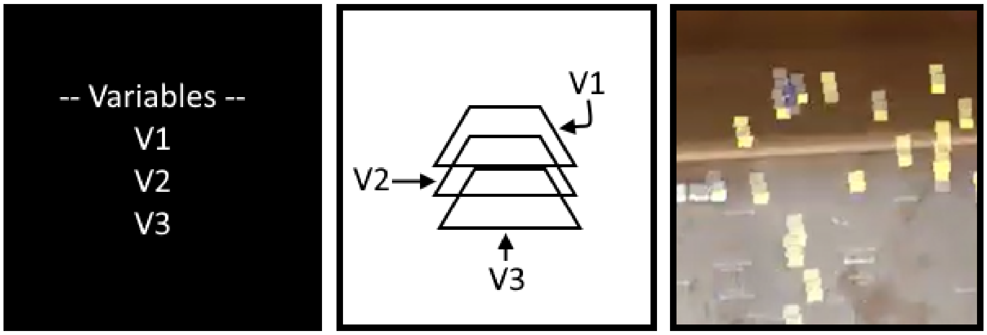


Figure 4.10: **LEFT:** The legend available in AR. **MIDDLE:** A diagram showing how the glyphs would be arranged; the user does not see this diagram. **RIGHT:** The glyphs as they appear in AR.

Figure 4.11: The colourmap for the pre-fit stage. Blue (RGB 0, 0, 127) represents the minimum and yellow (RGB 255, 255, 127) represents the maximum.

The AR interface presents the glyphs to the user. In the pre-fit stage, the glyphs represent the values of the data. We use glyph-based visualization to present the data at the most granular level, so the user can avoid committing the *ecological fallacy*. The ecological fallacy is a type of logical fallacy where aggregated information is used to predict individuals [Salkeld and Antolin, 2020]. For instance, if the average level of lake pollution is low, we commit the ecological fallacy by claiming that there are no polluted lakes in the region using the average. By allowing the user to examine data on a per-point basis, all possible trends become apparent to the user.

The ordering of the glyphs in the Variable Picker determines the level of the glyphs in the composite. For instance, if the selected variables are *V1*, *V2*, and *V3* in the Variable Picker, then each glyph composite has three glyphs. The top glyph is *V1*, the middle one is *V2*, and the bottom one is *V3*. Close to the tablet, the user can view the AR-based legend (Fig. 4.10-LEFT). To assign a colour to each glyph, we use the colourmap in Fig. 4.11. The glyph uses colour to convey normalized values between zero and one. The normalization is at the individual level of a variable–i.e., let v be the normalized value, x_i being the original value, and V be the set of numbers for a variable, then $v = \frac{x_i - \min V}{\max V - \min V}$. The positions of the glyph composites are based on the geospatial positions (i.e. latitude and longitude). The current version of Gander treat discrete variables in same way with the continuous ones; both types of variables are normalized in the same manner.

An important task is to find if the variables are correlated or not. For two variables to be correlated, they must meet one of the two requirements:

- **Positive correlation:** If one variable is increasing in value, so is the other one.
- **Negative correlation:** If one variable is increasing in value, the other one's value is decreasing.

If there are more than two variables, the user performs pairwise comparisons among the variables. Including correlated variables in a model goes against the goal of

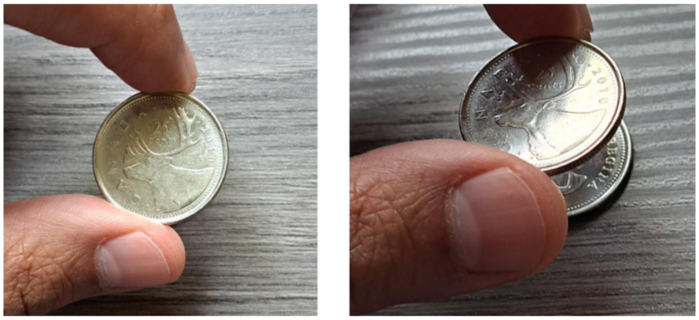


Figure 4.12: An example of the parallax effect with two Canadian quarter coins. **LEFT:** The bottom coin is occluded. **RIGHT:** After changing the camera's angle, the bottom coin is now visible.

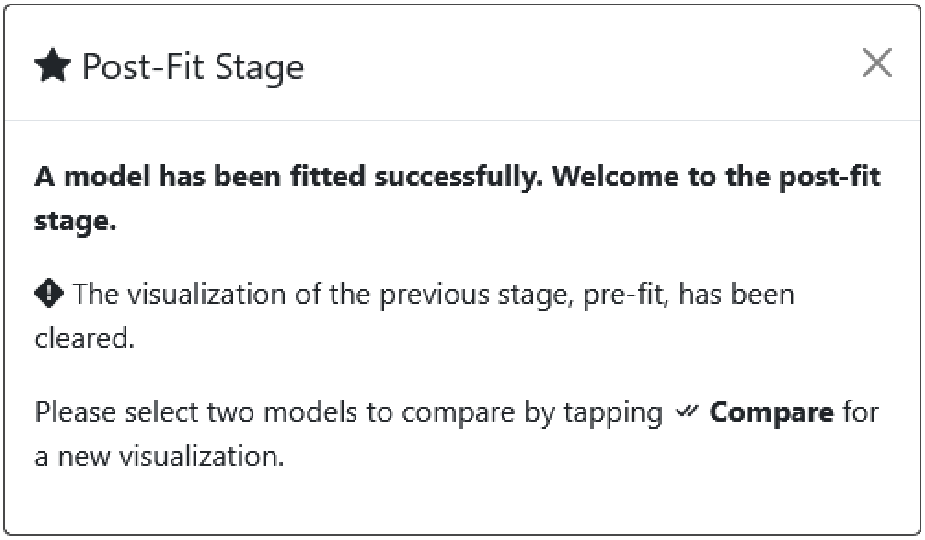
parsimony indicated in Section 2.1). Therefore, if two or more variables are similar to each other, they should consider removing them.

Another important task is for the user to find if the data tend to be distributed in the same way throughout the map. For instance, one area on the map might have a relatively elevated level than the other areas. The differences in levels may indicate the presence of spatial autocorrelation. The visual inspection is not sufficient, however. They must confirm this using other techniques.

The user can use the parallax effect (see Rouan [2015], Fig. 4.12) to decomposes the glyphs. By changing their viewing angle (e.g. tilting their head left and right), they can align the glyphs together without making any input to the tablet or to the OST-HWD. Where two glyphs intersect, their values are multiplied together. For instance, if Glyph A has the value of 0.5, and Glyph B has the same value, the overlapping value then has the value of 0.25. By showing the multiplication of the glyphs together, the user should be able to indicate if there is any multiplicative effect between the selected variables.

4.2.3 Interaction Stage 3: Post-fit

After fitting a model, Gander displays the post-fit alert dialog (Fig. 4.13). The alert box tells the user that the pre-fit stage has ended, and the original glyphs from the stages have been removed. In the post-fit stage, the user now compares the likelihood effect sizes of the model against another one. The other model, considered a full model, should have more variables than the selected one. The effect size is computed as:



★ Post-Fit Stage ×

A model has been fitted successfully. Welcome to the post-fit stage.

◆ The visualization of the previous stage, pre-fit, has been cleared.

Please select two models to compare by tapping ✓ **Compare** for a new visualization.

Figure 4.13: The post-fit alert dialog.

$$E_L = \frac{p_i}{\max\{p, q\}}$$

where $E_L \in [0, 1]$ represents a goodness of fit, p_i represents likelihood of a model, $\max\{p, q\}$ represents the maximum likelihood value of both models. A higher E_L indicates a better likelihood. We create this effect size based on the work by Johnston et al. [2006] which involves normalization of likelihood ratios. To prevent underflowing (i.e. values close to zero being treated as zero by a computer), we used log-likelihoods in our implementation of E_L instead.

Our implementation of E_L is for a specific effect size instead of the whole model, meaning that each glyph represents a goodness-of-fit of a model solely for the data point. This is to encourage the user to explore likelihoods as if they were the data like in the pre-fit stage. Exploring likelihood this way may help the user to further diagnose spatial autocorrelation. To our knowledge, trying to identify spatial autocorrelation in likelihood is a novel method in spatial analysis. The method that is most similar to this one is GWR where hierarchical local models are created. However, in GWR, the goodness-of-fit measure is still done at the global level–i.e., we do not evaluate the effectiveness of the local models.

Sections 4.2.3, 4.2.3 describe the other changes from the pre-fit, to the post-fit stage.

Tablet Interface

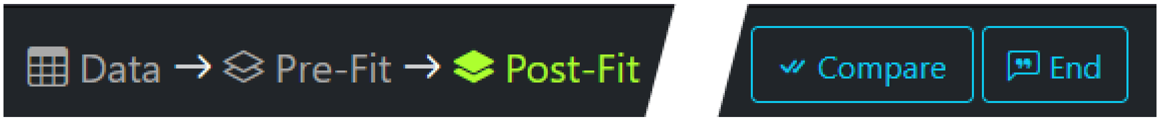
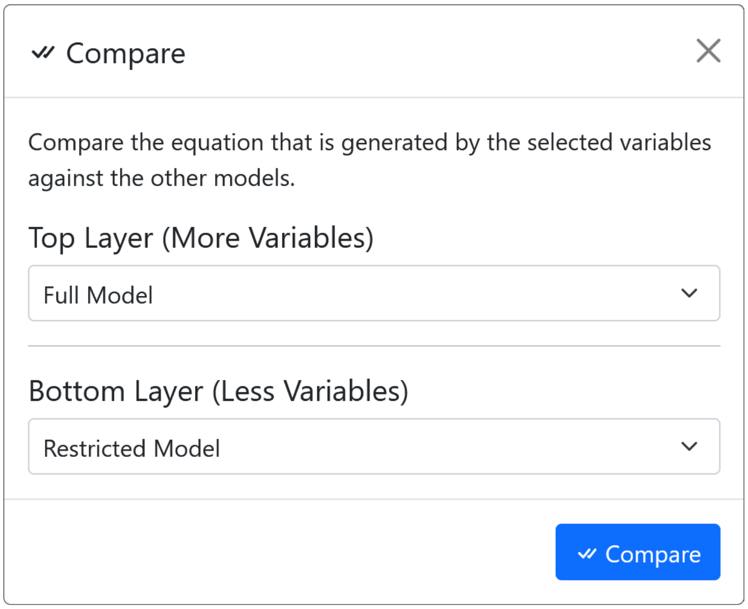


Figure 4.14: The new post-fit menu. The empty middle part has been removed for better spacing.



Compare the equation that is generated by the selected variables against the other models.

Top Layer (More Variables)

Full Model

Bottom Layer (Less Variables)

Restricted Model

Figure 4.15: The Model Comparer.

The map interface is largely the same as in the pre-fit stage, with two differences. First, Gander displays a dialogue showing that the pre-fit glyphs have been cleared and warns the user that they are now comparing likelihoods (Fig. 4.13). Secondly, a new menu option, "Compare", is present instead (Fig. 4.14). Tapping on "Compare" launches the Model Comparer (Fig. 4.15), allowing the user to compare two models in terms of their E_L, and whether they have any spatial differences (i.e. spatial autocorrelation). This essentially allows them to perform visual likelihood ratio tests.

The navigation bar also contains a button with the label "End." This button terminates both the tablet-based Gander and the AR-based Gander.

Augmented Reality Interface

The glyph-based visualization is similar to the previous stage, with a purple line representing the map boundary (Fig. 5.1-C). However, each glyph represents E_L instead. The overlapping glyphs represent differences of E_L between the two models. To indicate that the user is comparing two models instead of one, Gander changes its colourmap to a red-cyan one (Fig. 4.16).

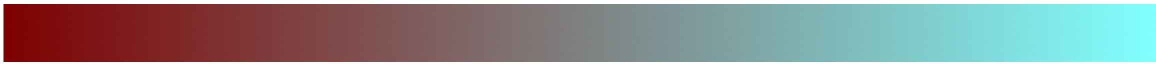


Figure 4.16: The colourmap used in the post-fit stage visualization. Dark red represents the minimum (0) and the cyan represents the maximum (1).

The user can again overlap the glyphs using the parallax effect. The overlapping areas display the absolute difference between the two models using the subtractive blend. Zero means no difference and one means maximum difference. The overlapping areas can help the user to understand the absolute differences between the goodness-of-fit of the two models. While the subtractive blend can be applied to as many glyphs as possible, since the subtractive blend is only comprehensible for two glyphs, we only allow the user to compare two models at a time.

4.3 Implementation

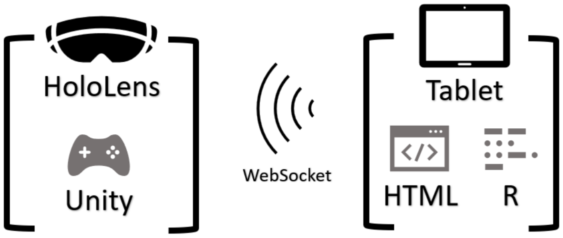
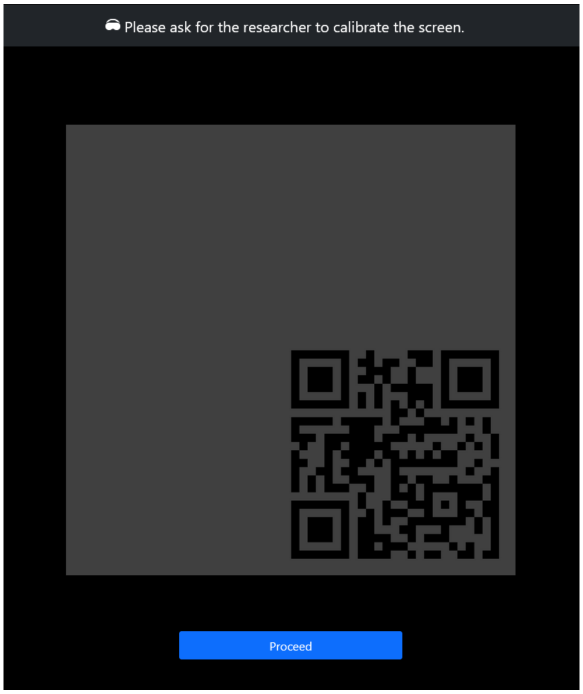


Figure 4.17: The "technology stack" diagram of Gander showing how the OST-HWD communicates with the tablet.

Gander consists of two pieces of software that run concurrently: the tablet-based software, and the AR-based software. The tablet-based software's primary task is to accept inputs from the user, and to communicate the input as WebSocket messages to AR-based software. Upon receiving the messages, the AR-based software interprets the messages and executes actions. Fig. 4.17 shows how devices are communicating with each other.

4.3.1 Tablet

The tablet-based software executes from a Node.js server. The server uses WebSocket to send the tablet's messages to the AR headset. We use Mozilla Firefox [Mozilla, 2023] to view the HTML content served from the server. For the web-based interface, Gander uses Bootstrap [Bootstrap, 2023], jQuery [OpenJS Foundation and JS Query Contributors, 2023], and pagePiling.js [Trigo, 2023]. When the user selects a map, Leaflet.js [V. Agafonkin and Leaflet maintainers, 2023] downloads the map data from Mapbox [Mapbox, 2022] and renders it on the screen during the pre-fit and the post-fit stages. Leaflet.js also supports map panning with touch gestures, and translation of touch input data to the AR-based software. SortableJS [K. Lebedev, et al., 2022] supports the drag-and-drop interactions in the dialog boxes. We used code by Palén [2012] to populate the dialog boxes.



Please ask for the researcher to calibrate the screen.

Figure 4.18: QR Code calibration system used in Gander.

Prior to the map selection, the tablet displays a quick response (QR) code for position registration. This step is crucial for ensuring that the AR content will align properly with the tablet. The user must properly align the OST-HWD's camera to the QR code. Once calibrated, the user is free to proceed. Fig. 4.18 shows our implementation of the calibration. Tönnis et al. [2013] calls this process "registration" and using QR code to register. Using a QR code to register content has been used in various projects (e.g. Rekimoto and Ayatsuka [2000], Satriadi et al. [2022]), and is supported on some OST-HWDs like Microsoft HoloLens v2 [Wen et al., 2021].

After selecting a map, the tablet reads the metafile which is a JSON file containing boundary information, and the map scaling information. The tablet applies the metafile's information to its own map display by changing the Leaftlet.js settings. It then messages the AR interface with a set of vectors obtained from converting boundary information in the metafile. The AR interface draws the purple boundary using the vectors that it receives.

When the user changes variables in the Variable Picker, the Equation Modeller, and the Model Comparer, the tablet-based software sends messages to the AR-based software to clear the AR glyphs, and redraw the glyphs based on the user's input. When transitioning from the pre-fit to the post-fit stage, the AR-based software clears the AR environment, changes the AR-based legend, and draws new post-fit glyphs.

To compute likelihood information for the post-fit, the server calls a R script file with the proper parameters. The R script then returns E_L values which the tablet can then send to the OST-HWD for drawing. Finally, the user can terminate the program with a button. The button sends a termination signal to the AR software as well.

Since the tablet-based software is web-based, it can be deployed on any tablet. During the studies, we used Microsoft Surface Book 3 (15 in. screen).

4.3.2 Augmented Reality

The AR-based software's functionality is to listen to the WebSocket messages, draw the map boundary, and manipulate the glyphs based on the user's input on the tablet. The software can remove glyphs, show glyphs, and rearrange the glyph composites. When the user pans the map through the tablet, the tablet emits transformed touch coordinates to the AR device. The device then moves the glyph and the boundaries based on the transformed coordinates. When the software receives the termination signal from the tablet, it also terminates itself. This is to reduce the chance of having multiple Gander instances running in AR at the same time.

We use Unity [Unity, 2023] with MRTK 2 [Microsoft, 2022] to implement the AR software. For the device, we use Microsoft HoloLens v2 which is a premium OST-HWD. NativeWebSocket [Drewyer, 2023] is used to support WebSocket listening on the HoloLens.

4.4 Colourmap Implementation

In Gander, we used two colourmaps which we implemented using our blending techniques. This section describes how we implemented them using shader-based programming. Furthermore, we also present colourmaps that used the shader-based technique, including ones not used in our research. A shader programming involves manipulating the pixels and the vertices of the objects that are about to be displayed on the screen. Because it is executed right before the display and only to the content visible to the user, shader-based programming is fast.

The simplest shader-based colourmap is the grey scale colourmap. It is simply a fragment shader that returns $RGB : (v, v, v)$ where $v \in [0, 1]$ is the normalized value of the object. However, this colourmap is not isoluminant. This means as v tends toward 0, the object will become more transparent and less visible on an OST-HWD. It may also be too simple for many use cases. We can make the colourmap more colourful with an additional shader pass. In the second pass, we force one or more colour channels to have fixed values through the use of a colour mask. A colour mask is a shader instruction to not alter the masked colour channels [Unity, 2023]. This means whatever colours in the masked channels in the first pass are unaffected by the second pass. A pseudo-code for this is available in Algorithm 4.4.1. In Fig 4.19, we provide examples of the colourmaps that we create using an additional pass. We found that some of our colourmaps bear some resemblance to the ones that researchers consider ideal like Cividis [Nuñez et al., 2018]. Unfortunately, if we want to implement more advanced shader operations like multiplicative blending, we must use a three pass-shader which we describe in Algorithm 4.4.1. This shader requires the use of the stencil buffer which can be resource-intensive.

4.4.1 Algorithms

No-blend of Simple Blend-Support Algorithm

If we do not need to use a blending operation or additive/subtractive blending, we can use a two-pass shader that performs the following:

Pass 1: Value-Assignment. The fragment shader function returns the following colour: $(v, v, v, 1)$ where $v \in [0, 1]$. v represents the value of the glyph. If blending is used, it is also performed in this pass.

Pass 2: Fixing the Channels. Add a colour mask to one or more channels. Assuming that we add the colour mask to R channel, the fragment shader returns $(0.5, 0, 0, 1)$. In this case, G, B still convey v from the values of v. We call this pass "fixing channels."

Blend-supporting Algorithm

If we require a blending operation, we may need a three-pass shader. This is more resource-intensive. For certain game engines like Unity, we need to enable certain settings like a 24-bit depth buffer. The outline of the three-pass shader is as follows:

Pass 1: Stencil Buffer. Check the stencil buffer. If the buffer for the pixel is not written, then the fragment shader returns $(1, 1, 1, 1)$. Otherwise, do not draw. This layer prevents over-blending of the same pixel. **This pass is only needed for multiplicative blend.**

Pass 2: Value-Assignment and Blend. The fragment shader function returns the following colour: $(v, v, v, 1)$ where $v \in [0, 1]$. v represents the value of the glyph. Also, perform the blend operation here.

Pass 3: Fixing the Channels. This is the "fixing channels" pass. Add a colour mask to one or more channels. Assuming that we add the colour mask to the R channel, the fragment shader returns $(0.5, 0, 0, 1)$. In this case, G, B still convey v from the values of v. Transparency (the alpha channel) should also be fixed to the desired level.

4.4.2 Sample Colourmaps Implemented with Shaders

Below are the examples of colourmaps that we implement using the shader-based methods. Our thesis work only uses Chicago, Ukraine, and Saga. We present other colourmaps to show what our shader-based technique can achieve.

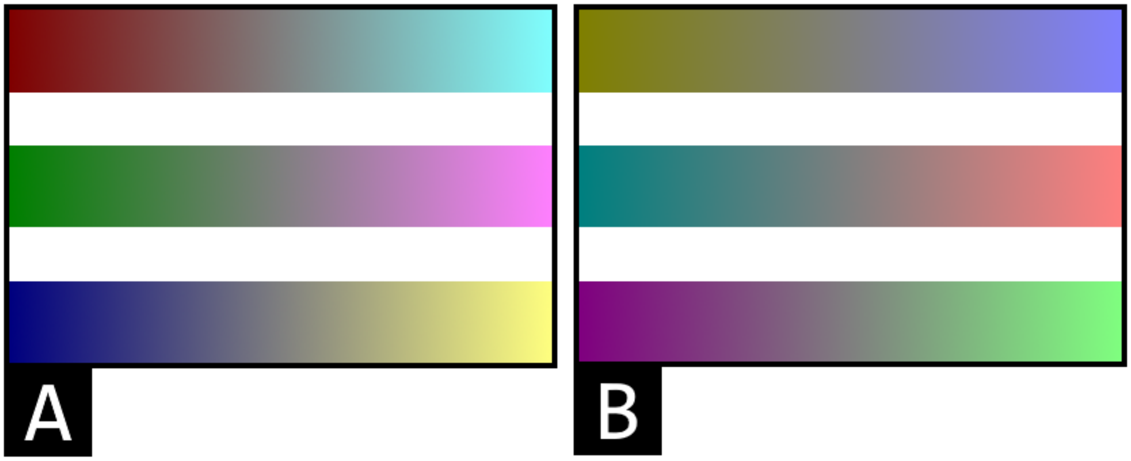


Figure 4.19: Example colourmaps. **A: TOP:** Chicago. **A: MIDDLE:** Newfoundland. **A: BOTTOM:** Ukraine. **B: TOP:** Pattalung. **B: MIDDLE:** Saga. **B: BOTTOM:** Adjuntas.

Fixing One Channel to 0.5

- **Chicago (Fig. 4.19-A:TOP):** We create this colourmap by fixing the red channel to 0.5. The lowest value of the map is $RGB : (0.5, 0, 0)$ and the highest value is $RGB : (0.5, 1, 1)$. We name this colourmap Chicago since it contains red and cyan, the colours in the flag of Chicago[1]. This colourmap is somewhat similar to the red-brown divergent colourmaps found in ArcGIS Esri Color Ramps [ArcGIS, 2022]. We chose this colourmap for the post-fit stage, and the participants of the synoptic study relied on this colourmap for the post-fit tasks.

- **Newfoundland (Fig. 4.19-A:MIDDLE):** We fix the green channel to 0.5. The lowest value of the map is $RGB : (0, 0.5, 0)$ and the highest value is

[1] Image of the Flag of Chicago: `https://design.chicago.gov/flag`

$RGB : (1, 0.5, 1)$. We name the colourmap Newfoundland after the Newfoundland Tricolour flag[2] which contains green and pink. This colourmap is similar to: (1) Harrower and Brewer [Harrower and Brewer, 2003]'s green to pink colourmap, and (2) Green and Pink 1 found in ArcGIS Esri Color Ramps [ArcGIS, 2022]. In one of our pilot evaluations[3], we found this colourmap is prone to colour distortion when using Microsoft HoloLens v2. In our experience, we find that the device has a tendency to introduce pink distortion.

- **Ukraine 4.19-A:BOTTOM):** We fix the green channel to 0.5. The lowest value of the map is $RGB : (0, 0, 0.5)$ and the highest value is $RGB : (1, 1, 0.5)$. We name the colourmap Ukraine after the flag of Ukraine[4]. This colourmap somewhat resembles Cividis Nuñez et al. [2018] which potentially makes it the most CVD-friendly among all sample colourmaps. We chose this colourmap for the pre-fit stage in the prototype of Gander. The participants of the synoptic study also interacted with the colour during the pre-fit tasks.

Fixing Two Channels to 0.5

- **Pattalung (Fig. 4.19-B:TOP):** We create this colourmap by fixing the red and green channels to 0.5. The lowest value of the map is $RGB : (0.5, 0.5, 0)$ and the highest value is $RGB : (0.5, 0.5, 1)$. We named the colourmap Pattalung, because the flag of Pattalung, Thailand[5] contains gold and purple.

- **Saga (Fig. 4.19-B:MIDDLE):** We fix the green and blue channels to 0.5. The lowest value of the map is $RGB : (0, 0.5, 0.5)$ and the highest value is $RGB : (1, 0.5, 0.5)$. We name the colourmap Saga after the flag of Saga, Japan[6]. This colourmap is similar to Conifer Forest found in ArcGIS Esri Color Ramps [ArcGIS, 2022]. We chose this colourmap for the elementary study (Chapter 6) because it is highly isoluminant. Further explanation on the benefit of isoluminance is provided in that chapter.

[2] Image of the Newfoundland Tricolour flag: `https://www.heritage.nf.ca/articles/society/newfoundland-republic-flag.php`

[3] Citation removed due to the manuscript being under review

[4] Image of the flag of Ukraine: `https://ukraine.ua/stories/ukrainian-flag-day/`

[5] Image of the flag of Pattalung: `http://www.phatthalung.go.th/2022/content/flag`

[6] Image of the flag of Saga: `https://www.crwflags.com/fotw/flags/jp-41.html`

- **Adjuntas (Fig. 4.19-B:BOTTOM):** We fix the red and blue channels to 0.5. The lowest value of the map is $RGB : (0.5, 0, 0.5)$ and the highest value is $RGB : (0.5, 1, 0.5)$. We name the colourmap Adjuntas, after the flag of Adjuntas, Puerto Rico[7]. The colourmap is similar to the purple and green colourmaps found in ArcGIS Esri Color Ramps [ArcGIS, 2022].

4.5 Design Limitations

Most of Gander's inherent design limitations are discussed in Chapter 7 and Chapter 9. In Chapter 7, we describe a walkthrough demonstration study that we conducted with multiple experts. We analyzed the feedback to identify multiple design issues (e.g., glyph-based visualization must be combined with other overview information, more interactivity). Then, we made several proposals to improve Gander. In Chapter 9, we provide a set of low-fidelity designs to improve Gander based on the participant feedback from the walkthrough demonstration study. In addition to the features suggested by the participants, extra features are proposed to help expand the vertical slice.

4.6 Modified Versions of Gander

The synoptic and the elementary studies used software derived from this software. Chapter 5 describes the version used in the synoptic study. Unlike the version of Gander described here, the version in Chapter 5 restricted the participants' actions to maintain a degree of experimental control over the study. That version also contains an online portal for conducting studies. The researchers could use the portal to track a participant's progress and to open online questionnaires for data collection. Hence, we can treat that version as the "kiosk-mode" version. Chapter 6 describes another research software derived from the Gander code base. Unlike the version described in this chapter and the "kiosk-mode" version, this version was so heavily stripped down that it could no longer be considered as a variant of Gander.

[7] Image of the flag of Adjuntas: `https://www.crwflags.com/fotw/flags/pr-aj.html`

4.7 User Scenario

Since Gander has a very complicated interface, we include a user scenario that outlines how someone may benefit from Gander. The user scenario is futuristic, because the current technologies, at the time of this writing, are unable to perform all the tasks described in this section.

4.7.1 Personas

Ariya Rungreong

Ariya is an analyst working at the Ministry of Environment. She is a recent graduate of a data science program. While not an enthusiast of MR, she is somewhat familiar with it because her brother likes to play co-operative VR games. She works with ecologists who often go out and collect samples of lake water. This means, Ariya rarely spends time in her office. To ensure that she is still able to perform statistical tasks, the ministry assigns her a small tablet and an AR headset. She uses the tablet for statistical analysis, and AR to extend the tablet's screen. The other ecologists also have similar sets of hardware which allow them to work collaboratively with Ariya.

Yuuto Mizumoto

Mizumoto is a senior ecologist working at the Ministry of Environment. Since he has recently injured himself while bicycling at Kejimkujik Park, he must recover at home. This means, he is unable to accompany Ariya and her colleagues. Normally, he only relies on his desktop computers to complete his tasks. However, there are occasions that he needs to use AR. For instance, the information charts may be in 3D which requires him to manipulate using hand gestures.

4.7.2 Story

Ariya and the ecologists at the Ministry of Environment have been collecting samples from various lakes around the Province. Before they close the project, they must first create machine learning models, and somehow communicate the models to Mizumoto. Since Mizumoto is unable to move, Ariya and her colleagues decide to have a meeting in AR.

In AR, Ariya and the ecologists use Gander to present the lake chemical data as glyphs. Each layer represents different types of chemicals. Mizuoto is very interested in the lake chemistry data, and how they may be connected to the recent algae bloom. He proposes that Ariya and her team creating models that can predict algae coverage of the lakes based on their lake chemistry. They launch Gander and start using it collaboratively.

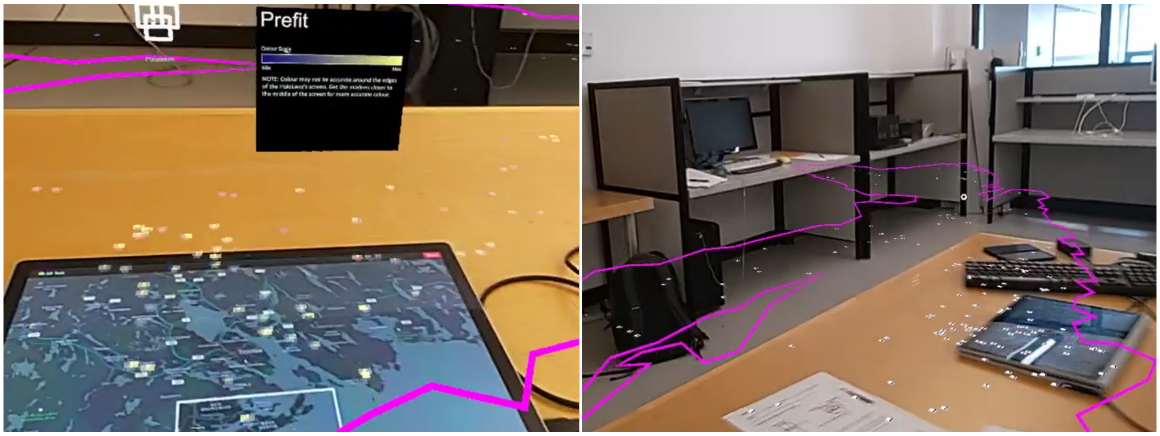
In the pre-fit stage, Ariya tries to find if the chemical volumes are correlated in any way. She believes that she should add sodium and chloride together as predictors. However, her colleagues state that these two chemicals tend to co-occur as salt, so she might be able to remove one of the predictors. They also find that the lakes in the northern region of the Province to have slightly different chemistry, so they decide to the region as a predictor.

Ariya fitted the model, and first compares it with the full model in the post-fit stage. She finds the improvement to be minimal. Mizumoto, therefore, asks her to try a slightly different model. She follows his recommendation, but the improvement is still quite small. She and her colleagues then try other models. Finally, they conclude that it is impossible to predict algae coverage using the data that they have. Mizumoto believes that more data are necessary. He decides to extend the data collection period much to the chagrin of Ariya and the ecologists.

Ultimately, with the additional data, Ariya and her colleagues are able to create a model that satisfies Mizumoto. He then presents the model to the Ministry of Environment. He is able to argue that someone must have been illegally importing banned detergent into the Province.

Chapter 5

The Synoptic Study: Understanding Glyph Field Navigation



NOTE: Colour may not be accurate around the edges of the HoloLens's screen. Get the markers closer to the middle of the screen for more accurate colour.

Figure 5.1: **LEFT:** Screenshot of P17 completing the study task with Radial in the pre-fit condition. **RIGHT:** Screenshot of P8 walking around the table to view Polyline glyphs.

We conducted multiple user evaluations to better understand how the users of Gander complete the pre- and post-fit tasks. Each evaluation varies in terms of experimental control, so that we can better understand each aspect of usability. This evaluation, dubbed the synoptic study, aims to have a middle level of control. The participants had the freedom to scan the visualization generated by Gander as they saw fit. Furthermore, they could complete the sub-tasks (e.g. finding correlation in the pre-fit stage, and finding goodness-of-fit in the post-fit stage) in any order or concurrently. However, there were multiple constraints–for example, the map and the data were pre-selected, and the tasks were structured. The synoptic study's main focus in on trajectory data. Since Gander uses a tablet and AR simultaneously, the study's goal is to better understand how the user would use the combination of devices to scan room-sized visualizations (a.k.a glyph fields) to accomplish IML tasks. We argue that this study aims to fulfill all aspects of the research objects (Obj1, Obj2, Obj3). For Obj1, We made a distinction between pre-fit and post-fit tasks during the study. For Obj2, we compared Radial (Fig. 5.4, also described in Sec. 5.2.2) against an existing control technique. For Obj3, we measured and compared the trajectory data that arose during the use of AR and a tablet. We call this study “the synoptic study” based on Andrienko and Andrienko [2005]’s task classification. A synoptic task with glyph visualizations involves the participants synthesizing information using multiple glyphs.

We collected and analyzed the trajectory data (gaze, position, and tablet-based scroll) generated by the participants while performing semi-naturalistic tasks with map-based data. We compared two glyph visualization techniques, the shape-based Polyline, and the colour-based Radial, to understand their influence on scanning behaviours. The control technique, Polyline by Opach et al. [2018], is a shape-based technique. It expresses multivariate values as zig-zagging lines cutting through squares, resulting in a glyph field consisting of many small line plots. We compared the technique against Radial. Unlike Polyline, Radial is a colour-based univariate technique, which arranges glyphs in a radial composite to express multivariate data. We designed Radial to support the detection of trends and outliers by scanning the glyph field. Additionally, when two or more Radial glyphs overlap, the colour of the overlapping area can express composite data (for example, by multiplying or subtracting overlapping values).

Twenty-four participants completed semi-naturalistic tasks based on geospatial regression analysis. The participants scanned the glyph fields using the AR+tablet interface displaying real-world map data. They then indicated the statistical information. Semi-realistic tasks yield observed behaviours that are more likely to have ecological validity than highly controlled and abstracted tasks, at least for the task domain. Our participants used the tablet primarily as an input device and to reference place names that weren’t displayed in AR. Polyline elicited more tablet-based panning of the glyph field–indicating that the technique led the participants to look at the individual glyphs more closely. Radial tended to elicit more gaze trajectories; the participants often examined the glyphs from afar, bringing them closer when reviewing the smaller overlapping regions of Radial glyphs. Despite having to scroll more, NASA-TLX and SUS scores favour the Polyline technique, possibly due to hardware-based colour distortions making some Radial glyph values hard to discern.

We argue that Gander follows the focus+context (F+C) paradigm; the tablet screen acts as the focus area due to it providing more visual information to the user while the AR provides larger but less detailed visualization [Baudisch et al., 2001]. However, we often observed the participants ignoring the tablet. This means, the participants did not use the tablet as the focus area. Thereby, Gander is not a true F+C system.

5.1 Multiple Linear Regression Steps

Multiple Linear Regression (MLR) is a technique that creates an equation that predicts a continuous dependent variable (DV). The equation is a linear combination of independent variables (IV). Although simpler than many other ML techniques, there are many considerations that one must make. In our study, we considered:

- **Multiplicativity:** We must consider if there is any interaction between the IVs. In MLR, a multiplicative term represents an interaction between one or more IVs [Braumoeller, 2004, Friedrich, 1982].

- **Parsimony:** A model must contain the least amount of IVs. If there are similar IVs, one of them should exclude the term. In MLR, lack of parsimony can sometimes manifest as multicollinearity which we can measure as Variance Inflation Factor (VIF) [Daoud, 2017]. We can also use techniques like Principle Component Analysis (PCA) to minimize it [Adnan et al., 2006].

- **Correlation:** The IVs must be able to explain the variances of observed DV values. Usually, we use measures like R^2 and adjusted R^2 to quantify this [Lewis-Beck and Skalaban, 1990].

- **Goodness-of-fit:** The values predicted by the equation must match the observed DV values. We can use likelihood to measure this [Johnston et al., 2006].

- **Spatial Autocorrelation:** Spatial data may have different distributions which may complicate fitting. Some measures like Moran's I [Moran, 1950] can describe the spatial autocorrelation. If the data are not spatial–i.e. do not have coordinates, spatial autocorrelation is not an issue.

There are other considerations that we ignore for the purpose of the study, because we believe they are ill-suited for our interface. For instance, we do not think that a glyph-based visualization can express the normality of the residuals.

5.2 Study Design

In our study, participants performed pre-fit and post-fit tasks relevant to multiple linear regression for geospatial data. We define a pre-fit task to be assessing variables before fitting a model. A post-fit task involves assessing models after fitting. Sec. 5.2.3 contains additional information on pre-fit and post-fit tasks. Our research questions are as follows:

RQ1: How do the techniques affect the user's scanning behaviours? Scanning is important for navigating and understanding glyph fields. We explored how each technique affected various types of trajectories possible within an AR+tablet interface. Polyline is likely good for identifying trends. Meanwhile, Radial, as a colour-based technique, may be good for pre-attentive perception. Unlike Polyline, Radial can express additional information which can complicate scanning behaviours. With different designs, these techniques should elicit different scanning behaviours as manifested in gaze, tablet-based panning, and OST-HWD position trajectory data. Additionally, the techniques can influence how the tablet can be used in conjunction with the tablet.

RQ2: What is the self-reported user experience of each technique? Since Polyline is shape-based and Radial is colour-based, we expect the user to have different experiences. We hypothesize that the user will have a better experience with Radial since Polyline requires the user to be closer in order to comprehend it. We administered self-reported questionnaires like System Usability Scale (SUS) [Brooke, 1996], and NASA-TLX [Hart, 2006]. Additional interviews supplement the questionnaires.

RQ3: What is the accuracy of each technique? We deployed a set of questionnaires to measure how well the participants understood the statistical information in the glyph field.

5.2.1 Participants

We recruited 24 participants (20 males, 4 females) using Dalhousie University's mailing lists. All participants were undergraduate and graduate students in the computer science program with some familiarity with MLR. One participant completed secondary education, 15 completed undergraduate studies, and eight possessed graduate degrees. Twenty-two participants had prior experience with mobile AR such as Pokémon Go. Eight had experience with OST-HWDs such as HoloLens. Seven had experience with virtual reality. The mixed reality experience can be overlapping–i.e. one participant could have experienced more than one technology. Each participant received 25 Canadian dollars for their participation.

5.2.2 Glyph Field

AR+Tablet Interface

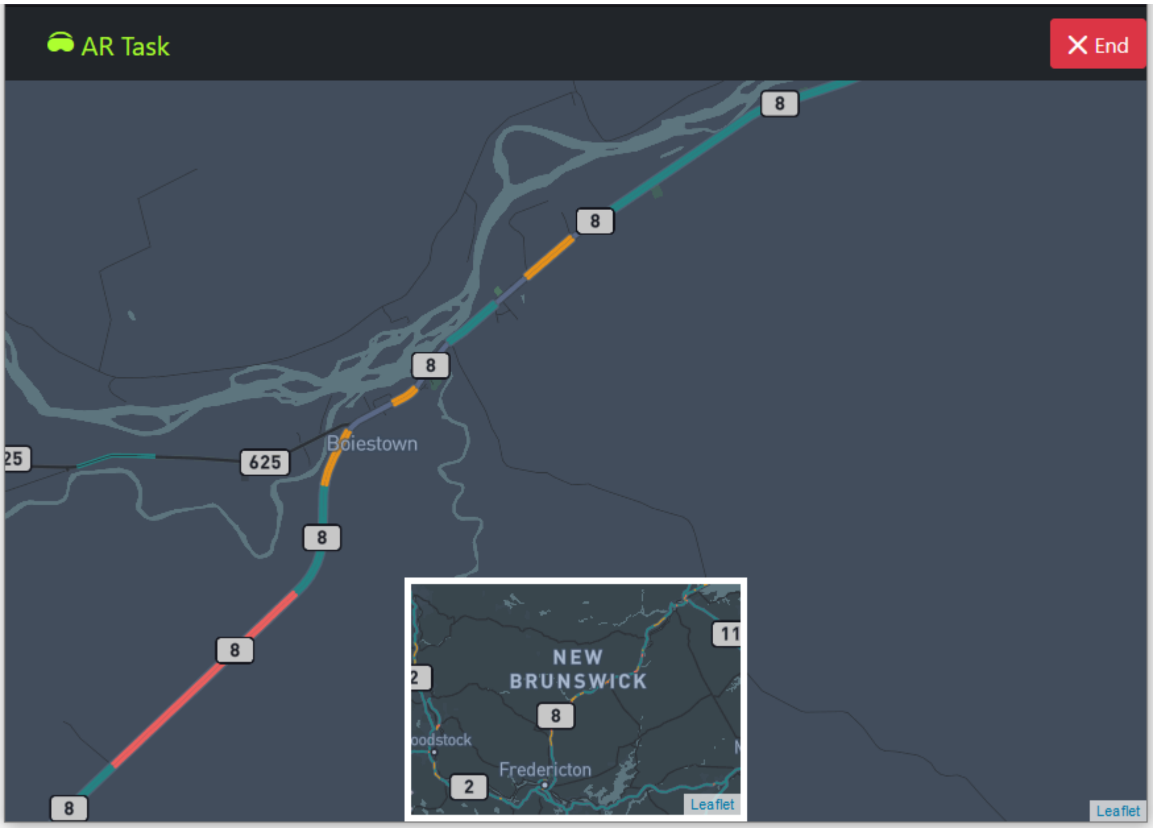
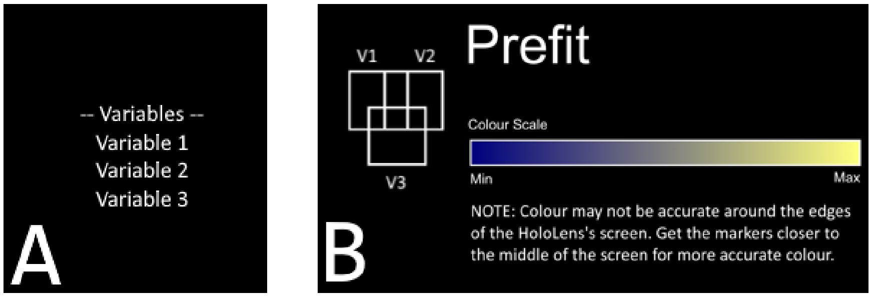


Figure 5.2: The tablet interface. The smaller inset is the virtual touchpad, which allows for rapid scrolling of the entire area. Tapping "End" means all sub-tasks have been completed.

An AR+tablet interface displayed a glyph field to the user. It consisted of a tablet (15-inch Microsoft Surface Book 3), laid horizontally, displaying a map, and the AR (Microsoft HoloLens v2) mounted a glyph field on the top. Glyph fields in the study



NOTE: Colour may not be accurate around the edges of the HoloLens's screen. Get the markers closer to the middle of the screen for more accurate colour.

Figure 5.3: The legends for pre-fit AR tasks (Post-fit versions were slightly different). **A:** The legend for Polyline. **B:** The legends for Radial, truncated here to fit on the figure.

were larger than the tablet which means most glyphs appear beyond the boundaries of the tablet's screen. The participants could navigate around the glyph field (i.e., standing up, or walking), or they could use a virtual touchpad (Fig. 5.2) on the tablet to pan the glyph field. We did not implement zooming, since it could act as a confounding factor in analyzing scanning behaviours.

Glyphs

The participants completed the tasks with Polyline and Radial. Fig. 5.6 shows our implementation of Polyline. Since the technique is a small multiple of the line chart, it contains an x-axis. The x-positions indicate variables in the pre-fit tasks and the models in the post-fit tasks. Since the glyph was too small to have x-labels, a list of variables/models was available in AR (Fig. 5.3-A) close to the tablet. The height indicates normalized values. During the pre-fit tasks, it indicates a value normalized to be between zero and one. In the post-fit tasks, it visualizes E_L an effect size based

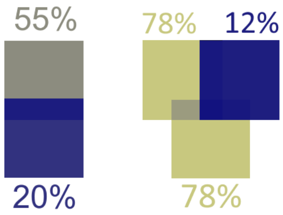


Figure 5.4: **LEFT:** Radial with two variables. **RIGHT:** Radial with 3 variables.

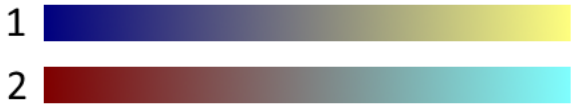


Figure 5.5: **1:** Colourmap for pre-fit tasks. **2:** Colourmap for post-fit tasks.

on the work of Johnston et al. Johnston et al. [2006]. The effect size formula is: $E_L = \frac{\ell_{i,j}}{\max \ell}$ where $\ell_{i,j}$ is a single MLR likelihood for the model i and the data point j. $\max \ell$ is the maximum likelihood for both models. E_L is bound between zero and one. The colour of the line indicated whether the tasks were pre-fit (blue) or post-fit (dark red). Polyline (Fig. 5.6) used zig-zagging lines cutting through squares to express multivariate values which makes it a shape-based technique.

Radial (Fig. 5.4) used colours to express its normalized values in the pre-fit tasks, and E_L in the post-fit tasks. The names of the colourmaps are Ukraine, and Chicago, and they are the same ones used in the prototype described in Ch. 4. For the implementation details, please refer to Section 4.4. Section 2.4 in Chapter 2 contains the colourmap design rationales. As a colour-based technique, Radial is suitable for pre-attentive tasks (i.e. before the user can pay attention to the glyphs [?]) and to provide overview information [Ropinski et al., 2011]. Since each glyph is univariate, we combined multiple glyphs into a single composite using a radial arrangement. The overlapping areas inside show multiplied values among glyphs during pre-fit tasks and absolute differences of the two models' E_L during post-fit tasks. An AR legend (Fig. 5.3-B) was available to the participants to help with understanding the arrangement. Unlike Polyline which is more difficult to compose, we created Radial with future extensibility in mind. For instance, Radial glyphs could be hovering on top of each other vertically to support 3D visualization. Since Polyline is a 2D technique, we limited our Radial arrangement to 2D in this study.

As Fuchs et al. [2017] indicated, there are many ways to arrange a glyph field. Since our study tasks are based on map-based data exploration, we generated glyph positions using the data's latitude and longitude positions.

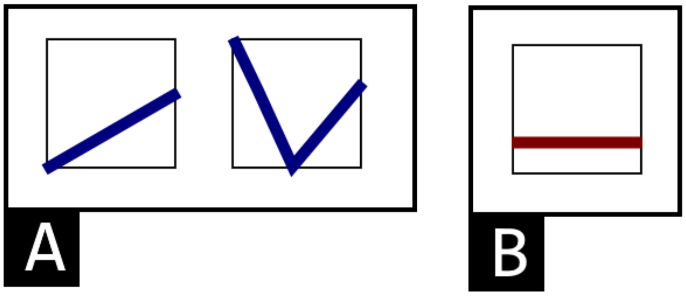


Figure 5.6: **A:** ***Left.*** Polyline for two-variable pre-fit tasks. ***Right.*** Polyline for three-variable pre-fit AR tasks. **B:** Polyline during Post-fit.

5.2.3 Protocol

To obtain semi-naturalistic results, our study tasks were based on exploratory geographic spatial analysis. In general, the participants scanned the glyph fields to obtain statistical information for a specific map and a specific set of variables/models. There were two maps: TO which is based on Toronto apartment scoring data [City of Toronto, 2021], and NS which is based on Nova Scotia lake chemistry data [The Government of Nova Scotia, 2021]. These maps represent different use cases: TO is a smaller map with urban data, and NS is a larger map with natural data.

At the beginning, all participants completed a demographic questionnaire. Then, they read a manual on the tasks and learned about the pre-fit and post-fit tasks. The pre-fit tasks were for setting up a model while the post-fit tasks were for assessing fitted models. We verbally quizzed the participants and informed them of what they needed to complete. To ensure the gaze tracking worked correctly, we asked each participant to perform an eye calibration for the OST-HWD. We then assigned them to one of the four groups based on their participant ID: G1, G2, G3, and G4. The study had a mixed design; each experienced all techniques and maps, but not all combinations of both. Please refer to Table 5.1 and Fig. 5.9 for more information about how the variables/models the participants worked with, and how these were mapped.

Once the participant was ready, they completed the following steps (also summarized in Fig. **??**):

1. AR-based training based on the assigned techniques (G1 and G4 with Polyline, and G2 and G3 with Radial). The training was for the pre-fit task. We did not specify to the participants how they should navigate the glyph field–e.g., we did not encourage nor discourage the participants from walking around the glyph field.
2. Pre-fit sub-tasks in AR for the assigned map and the assigned techniques. Each participant completed these IML actions while verbalizing their actions: (1) finding minimum and maximum values, (2) finding the correlation of the variables to assess parsimony, (3) finding the multiplicative effect between the values, and (4) finding spatial autocorrelation. Table 5.1 describes the variables that the participants analyzed. Table 5.1-A1 was for G1, G3. Table 5.1-A2 was

for G2, G4.

3. Completed Effect Size Questionnaire (ESQ, more information in Sec. 5.3.4) for the sub-tasks of the above step.

4. Completed NASA-TLX.

5. Repeated pre-fit sub-tasks with new variables. Table 5.1-B1 for G1, G3. Table 5.1-B2 for G2, G4.

6. Completed Effect Size Questionnaire for the sub-tasks of the above step.

7. Completed NASA-TLX.

8. Post-fit sub-tasks in AR for the assigned map and the assigned techniques. Due to the post-fit tasks being similar to the pre-fit task, no training was provided. Each participant completed these IML actions: (1) assessing the goodness-of-fit of both models, (2) finding the minimum and maximum goodness-of-fit, and (3) finding spatial autocorrelation. Table 5.1 describes the variables that the participants analyzed. Table 5.1-C1 was for G1, G3. Table 5.1-C2 was for G2, G4.

9. Completed Effect Size Questionnaire for the sub-tasks of the above step.

10. Completed NASA-TLX.

11. Completed SUS for the technique.

12. Completed a short semi-structured interview. The questions were: (1) What do you think about the overall interface?, (2) What do you think are the main benefits of the interface?, (3) How can the interface be improved?, (4) Did you experience any difficulty with the augmented reality device?.

13. Repeated the steps above (Steps 1-12) with a different technique. G1, G4 repeated with Radial and G2, G3 repeated with Polyline. The map also changed. Therefore, G1 and G2 repeated the tasks with NS using the variables and models in Table 5.1-A2, -B2, -C2. G3 and G4 repeated the task with TO with the variables and the models in Table 5.1-A1, -B1, -C1.

14. Compared both techniques in an exit interview. The questions were: (1) Which of the two techniques do you prefer?, (2) Can you tell me why?, (3) Which of

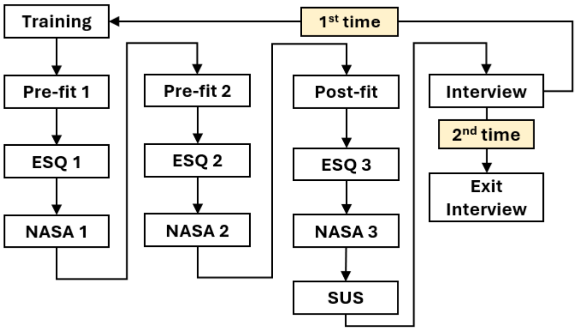
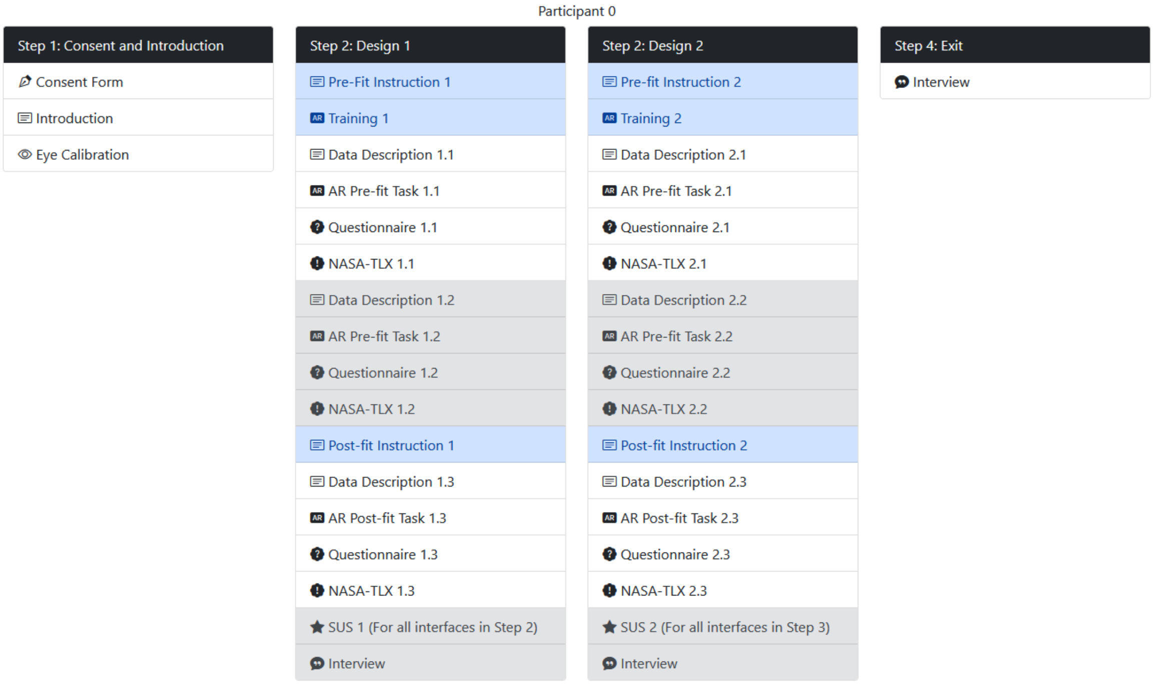


Figure 5.7: The flowchart for the procedure performed in the synoptic study.



Participant 0

Step 1: Consent and Introduction
- Consent Form
- Introduction
- Eye Calibration

Step 2: Design 1
- Pre-Fit Instruction 1
- Training 1
- Data Description 1.1
- AR Pre-fit Task 1.1
- Questionnaire 1.1
- NASA-TLX 1.1
- Data Description 1.2
- AR Pre-fit Task 1.2
- Questionnaire 1.2
- NASA-TLX 1.2
- Post-fit Instruction 1
- Data Description 1.3
- AR Post-fit Task 1.3
- Questionnaire 1.3
- NASA-TLX 1.3
- SUS 1 (For all interfaces in Step 2)
- Interview

Step 2: Design 2
- Pre-fit Instruction 2
- Training 2
- Data Description 2.1
- AR Pre-fit Task 2.1
- Questionnaire 2.1
- NASA-TLX 2.1
- Data Description 2.2
- AR Pre-fit Task 2.2
- Questionnaire 2.2
- NASA-TLX 2.2
- Post-fit Instruction 2
- Data Description 2.3
- AR Post-fit Task 2.3
- Questionnaire 2.3
- NASA-TLX 2.3
- SUS 2 (For all interfaces in Step 3)
- Interview

Step 4: Exit
- Interview

Figure 5.8: The online portal for the study.

the techniques are better for the following tasks–*Identifying correlation and interaction, Identifying differences between the models, Identifying regional trend on the map?*.

Each session was supposedly 90 minutes long. However, the participants often took up to two hours. Each participant received 25 Canadian dollars at the end of the study session. Since the study contained many steps, we implemented a HTML portal that performed the following: (1) serving as a checklist, (2) opening appropriate web-based forms (for ESQ, NASA-TLX, SUS), and (3) launching the AR task set-up. Fig. 5.8 contains the screenshot of the portal.

A1: Pre-fit/2 Variables/TO	**A2: Pre-fit/2 Variables/NS**
Security: 1-5 score representing security of a building	Calcium: Calcium level in a lake.
Stairwells: 1-5 score representing the quality of the stairwells.	Chloride: Chloride level in a lake.
B1: Pre-fit/3 Variables/TO	**B2: Pre-fit/3 Variables/NS**
Graffiti: 1-5 score indicating the presence of graffiti on the building. 5 means no graffiti.	Iron: Iron level in a lake.
Exterior Cladding: 1-5 score indicating the quality of the cladding for the building.	Manganese: Manganese level in a lake.
Exterior Ground: 1-5 score indicating the quality of the outside area around the apartment.	Potassium: Potassium level in a lake.
C1: Post-fit/TO	**C2: Post-fit/NS**
Model 1: Score = Graffiti + YearBuilt + Graffiti × YearBuilt	Model 1: TCU = Iron + Silica + Iron × Silica
Model 2: Score = Graffiti	Model 2: TCU = Iron

Table 5.1: The variables/models each participant interacted with in the pre-fit and post-fit tasks. TCU is "True Colour Unit", a score representing the colour of the water in a lake with particulate matter removed by centrifugation [Health Canada, 1995].

Maps and Variable/Model Mapping

Each map contains different sets of variables and models for the participants. Table 5.1 refers to the list of variables and models. Meanwhile, Fig. 5.9 shows how the variables and models are mapped in AR. The figure uses the same colourmap that the participants would use in AR.

5.3 Analysis and Results

We collected interview data, video data, self-reported measures (NASA-TLX, ESQ, SUS), and HoloLens log data. To answer RQ1, we analyzed the video data and three types of trajectories (gaze, touchpad scrolling, and HoloLens positions) extracted from the HoloLens log data. To answer RQ2, we performed a thematic analysis (TA) on the interview data, and quantitative analyses on the self-reported measures (NASA-TLX, SUS). To answer RQ3, we analyzed the ESQ data for the two techniques. We also compared participant ESQ scores against our own estimates (see Fig. 5.15). We used R libraries for quantitative analyses, and we set 0.05 to be the threshold for statistical significance for Type I Error.

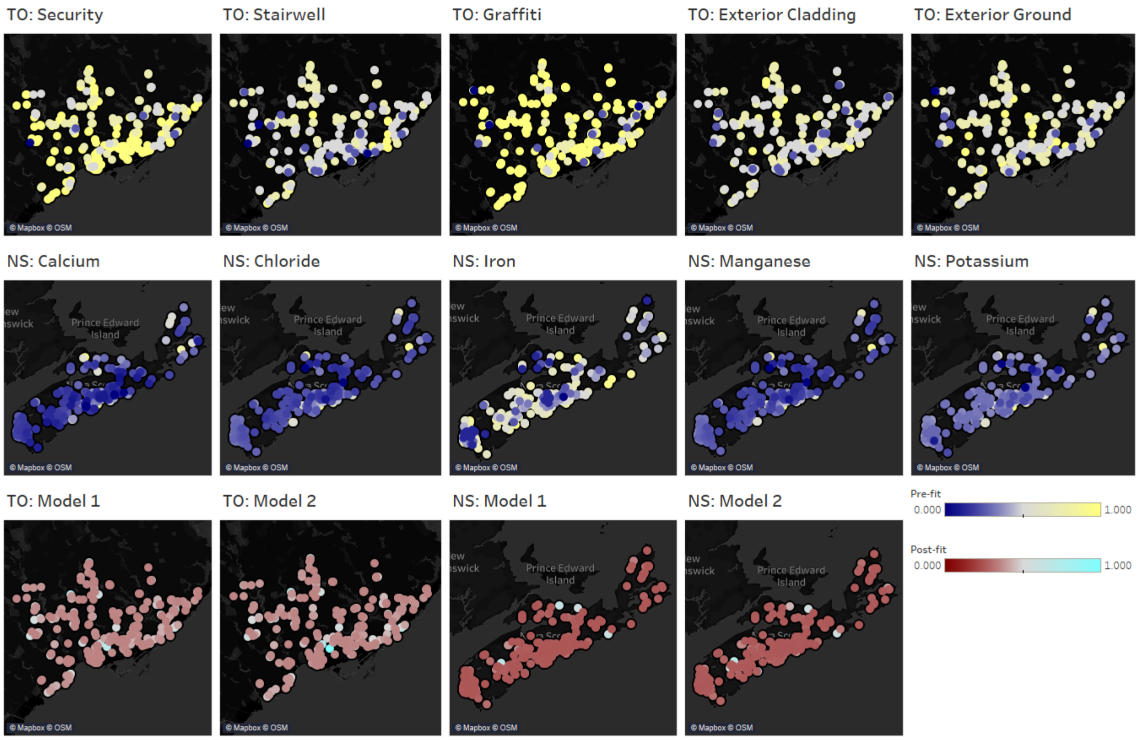


Figure 5.9: The small multiples of Toronto and Nova Scotia maps showing how the variables/models are mapped in AR.

5.3.1 Video Analysis

We conducted two passes on the video recordings, using `BORIS` [Friard and Gamba, 2016] video coding software. In the first pass, we flagged sequences that should not be considered in certain analyses, as detailed below. In the second pass, we coded participant actions to compare how tasks were completed using each technique and map. We then used `TraMineR` to process and to perform tree-regressions on the second-pass sequences to: (1) determine how participants' actions related to trajectories, and (2) how they completed the sub-tasks.

Using the "bottom-up" approach (see [Braun and Clarke, 2006]), we identified the following codes for participant actions:

Around Table (AT) The participant physically moved around the table.

Detach AR Screen (DARS) The participant scrolled the AR content such that the map became visually decoupled from the tablet. They then ignored the tablet's content. This means Gander no longer conformed with the F+C paradigm.

Explaining (E) The participant and/or the experiment facilitator were speaking. This could affect the participants' other actions. For instance, they may stop moving to talk to the facilitator. We found the interview data to be more insightful. Therefore, we did not analyze the utterance in detail–except for identifying the sub-tasks.

Looking Behind (LB) The participant was looking behind them, away from the tablet.

Looking off the Screen (LOS) The participants were not viewing the tablet.

Move Chair (MC) The participant moved their chair away from the table while staying seated. This code no longer applied when the participant returned.

Scrolling Touchpad (ST) The participant used the touchpad to pan the content.

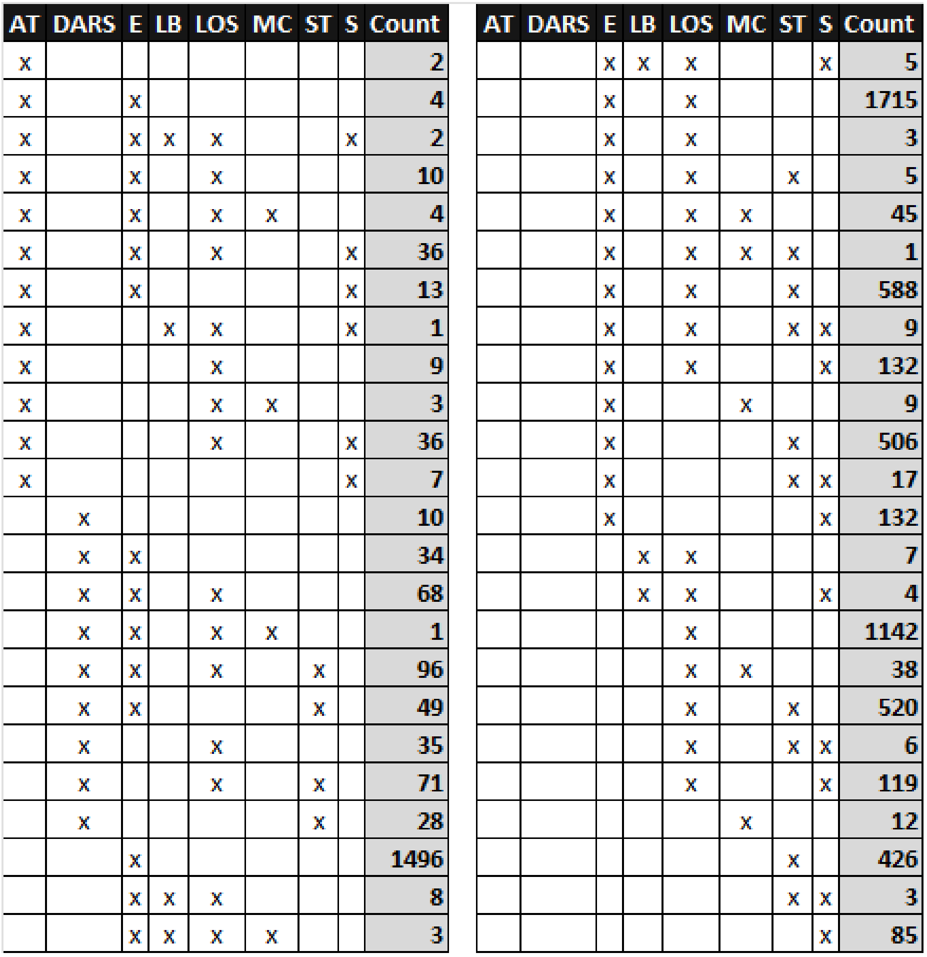
Standing (S) The participants were standing or not seated.

Some videos have codes for errors. The codes below were also used to indicate trajectories to exclude in the subsequent trajectory analyses:

Over-scrolling (OS) The participant lost the AR content by scrolling the touchpad too far. In the analysis of touchpad scroll trajectories, we removed AR data with this code.

Video Recording Failure (VRF) The HoloLens' video recording abruptly ended. We removed the following data from the sequence tree-regression with this code: P2 with Polyline and Pre-fit with 2 variables, P2 with Polyline and Pre-fit with three variables, P4 with Radial and Post-fit, P19 with Radial and Pre-fit with 3 variables, P21 with Polyline and Pre-fit with 3 variables.

Codes were sometimes combined. For instance, an action would be coded as S+ST if the participant was scrolling the touchpad while standing. Table 5.2 presents the frequencies of code combinations. We used `TraMineR` to perform a tree regression on sequences of these codes, using the methods outlined by Studer et al. [2011] to see if the study variables could affect the sequences of actions. The predictors were Technique, Map, and whether the tasks were pre-fit or post-fit. The regression shows



AT	DARS	E	LB	LOS	MC	ST	S	Count
x								2
x		x						4
x		x	x	x			x	2
x		x		x				10
x		x		x	x			4
x		x		x			x	36
x		x					x	13
x			x	x			x	1
x				x				9
x				x	x			3
x				x			x	36
x							x	7
	x							10
	x	x						34
	x	x		x				68
	x	x		x	x			1
	x	x		x		x		96
	x	x				x		49
	x			x				35
	x			x		x		71
	x					x		28
		x						1496
		x	x	x				8
		x	x	x	x			3
		x	x	x			x	5
		x		x				1715
		x		x				3
		x		x		x		5
		x		x	x			45
		x		x	x	x		1
		x		x		x		588
		x		x		x	x	9
		x		x			x	132
		x			x			9
		x				x		506
		x				x	x	17
		x					x	132
			x	x				7
			x	x			x	4
				x				1142
				x	x			38
				x		x		520
				x		x	x	6
				x			x	119
					x			12
						x		426
						x	x	3
							x	85

Table 5.2: Frequencies of the code combination.

differences in sequences between the task types (Pseudo-$F = 3.980, p = 0.001$; Levene's $W = 6.720, p = 0.006$), but the effect size is small (Pseudo-$R^2 = 0.030$). Since TraMineR determined that no more significant statistics could be computed, it applied early stopping to the regression. Therefore, additional statistics are unavailable. We can conclude that Technique, and Map do not have any discernible effect.

In addition to the analysis of the action codes, we have sub-task sequences which we inferred from the videos' subtitles. For the pre-fit videos, the sub-task codes are: (1) finding the minimum values, (2) finding the maximum values, (3) finding the correlation, (4) finding spatial autocorrelation, and (5) finding interaction. Our coding of sub-tasks allows for concurrent actions (e.g. finding the minimum and the maximum values at the same time). We performed a tree-based regression with Technique and Map being the predictors, and the sub-task sequences combined with

E are the response variables. We decided to add E since we found the conversations between the participants and the facilitator could affect the orders of the sub-tasks. The regression could not generate any model, and assigned the Pseudo-R^2 value of 0. For the post-fit videos, the sub-task codes are: (1) finding the minimum goodness-of-fit, (2) finding the maximum values, (3) finding the difference of the models, and (4) finding spatial autocorrelation. Another tree-based regression with Technique and Map being the predictor on the sub-task sequences combined with E as the dependent variable did not yield any model (Pseudo-$R^2 = 0$). Therefore, Technique and Map did not have any impact on the order of the sub-tasks'.

We found the participants almost always maintained the same forward heading. Those who turned back (P4, P8, P9, P12, P13, P16, P24) only did so temporarily, and were aware of the glyphs' real orientations. The participants often performed actions (62% of the actions per participant) without looking at the tablet–meaning the AR+tablet interface was not used as a F+C one. It is important to note that we did not instruct the participants on how to orient themselves.

5.3.2 Trajectory Analysis

To answer RQ1, and to understand the participants' scanning behaviour, we computed three types of mean-squared displacement (MSD): gaze, touchpad scroll (i.e. using the tablet to pan the AR content), and position. We computed MSD for each trajectory using following formula Poupard et al. [2019]: (2D) $MSD = Var(X) + Var(Y)$, and (3D) $MSD = Var(X) + Var(Y) + Var(Z)$, where X, Y, Z are coordinates of the positions. There are other measures than MSD, but they are not usable due to *random walk*, which is a tendency for someone to move randomly [Almeida et al., 2010, McLean and Skowron Volponi, 2018]. Table 5.3 describes the results of the trajectory tests. Since MSD is not a temporal measure, we performed another ANOVA on the task completion durations (millisecond). Since all effects have two levels each, a posthoc test is unnecessary. Although the trajectory data are not normally distributed, we report their means and standard deviations because we performed parametric ANOVAs.

	(a) MSD (Gaze)			**(b) MSD (Scroll)**			**(c) MSD (Position)**			**(d) Duration**		
Effects	χ^2	p	η^2	χ^2	p	η^2	χ^2	p	η^2	χ^2	p	η^2
Technique (T)	4.352	0.037*	0.028	6.619	0.010*	0.062	2.143	0.143	0.014	2.192	0.139	0.009
Map (M)	0.967	0.325	0.006	0.252	0.616	0.003	0.012	0.913	<0.001	0.602	0.438	0.002
Number of Variables or Models (N)	3.789	0.052	0.023	2.547	0.111	0.022	0.000	0.997	<0.001	0.362	0.547	0.082
Post-fit	x	x	x	x	x	x	x	x	x	87.042	<0.001*	0.335
T:M	1.276	0.259	0.034	0.648	0.421	0.008	5.676	0.017*	0.062	0.088	0.766	0.001
T:N	1.289	0.256	0.008	0.033	0.856	<0.001	0.832	0.552	0.002	3.346	0.067	0.013
M:N	0.061	0.805	<0.001	0.189	0.664	0.002	0.832	0.362	0.005	0.005	0.944	<0.001
T:M:N	0.582	0.446	0.004	0.092	0.762	0.001	1.078	0.299	0.007	1.400	0.237	0.005

Table 5.3: Combined ANOVA (Type II, Wald) tables made with log-linked Γ generalized linear models (GLM). The random effect was the participant, all effect had the df of 1. "x" means value unavailable due to fitting errors. Number of Variables or Models means whether the participants were analyzing two or three variables/models–independent of pre-fit or post-fit. Post-fit means whether the AR tasks are pre-fit or post-fit. * means $p \leq 0.05$.

Gaze Trajectory Analysis

Analyzing gaze data allows us to understand better how the participants looked at the visualization when completing the tasks. Before we can compute MSD, we must pre-process the data. We computed the point of intersection (PoI) between the participants' gaze rays and the horizontal plane with the tablet at the origin. Since the participants were not always looking at the visualization, some PoIs had extreme positions. We first filtered out any PoI generated when the participants were not looking downward. Then, we used `tclust` by Fritz et al. [2012] to trim the PoIs with the following parameters: $k = 3$, $\alpha = 0.05$, $iter.max = 100$. Fig. 5.10 shows the final results of the conversion process. Finally, we computed 2D MSDs per trial with the trimmed PoIs. Our ANOVA analysis (Table 5.3-a) shows that only Technique is statistically significant. The effect size is strong with Nakagawa's R^2_{GLMM}'s (ψ_1) being: $R^2_m = 0.113, R^2_c = 0.626$. The result shows the participants' gazes tended to travel further when using Mondrian–meaning that Mondrian ($\bar{x}_{MSD} = 0.217m^2, sd = 0.178m^2$) tended to encourage more gaze exploration than Polyline ($\bar{x}_{MSD} = 0.198m^2, sd = 0.174m^2$).

Radial also has one extra feature that Polyline does not: expression of multiplicative values in the pre-fit tasks and expression of log-likelihood difference in the post-fit tasks through the overlapping areas of the glyphs. We tested if the overlapping caused the participants to become more fixated. We define a new variable, Overlap, be the time when the participants were performing the sub-tasks which involved

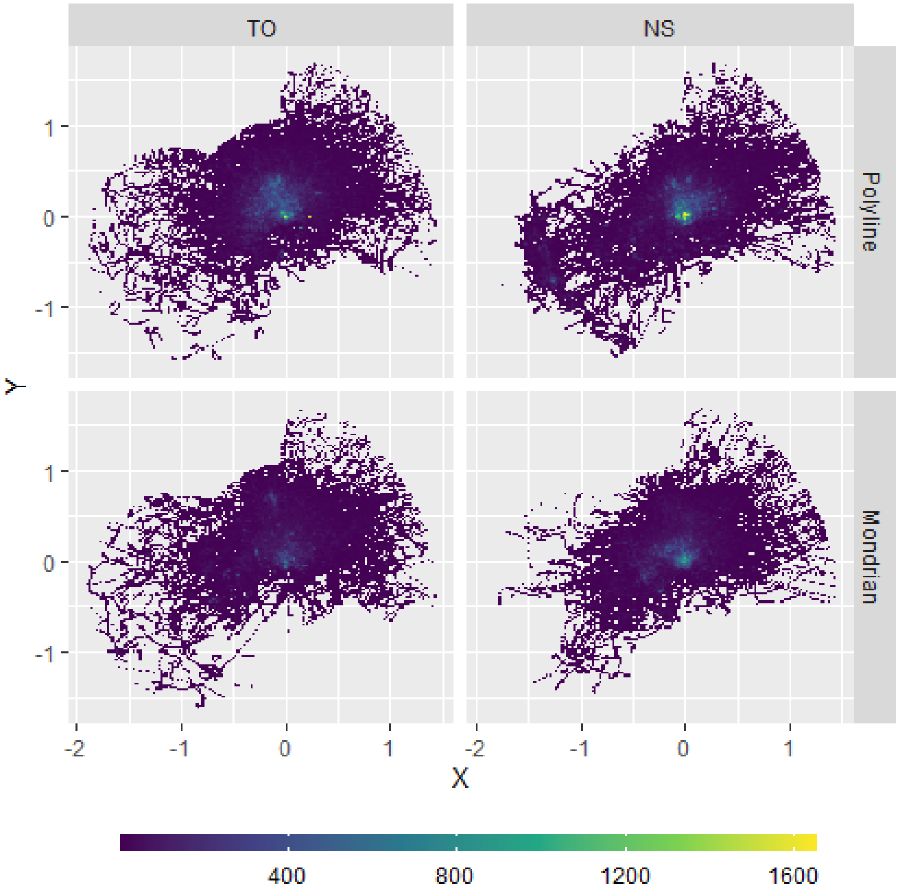


Figure 5.10: The heatmaps (2D histograms) of combined gaze PoIs grouped by Map and Technique. The Bin width is 0.05. X and Y's units are metres. Polyline histograms are brighter at the centres due to the participants focusing more around the origin. Radial histograms are darker since the gaze trajectories are more spread out.

looking at the overlapping areas between the Radial glyphs. Since we relied on the video tags to determine when Overlap occurred, we excluded gaze trajectories whose video recordings were tagged with VRF. We performed a Type II ANOVA (Wald's test) using a log-linked Γ generalized linear model (GLM) with MSD being the response values. The fixed effects are: Overlap ($\chi_1^2 = 4.773, p = 0.029, \eta^2 = 0.089$), Map ($\chi_1^2 = 0.859, p = 0.354, \eta^2 = 0.063$), and the interaction between the two ($\chi_1^2 = 0.036, p = 0.849, \eta^2 = 0.001$). The random effect is the participants. Nakagawa's R^2_{GLMM} (ψ_1) are: $R^2_m = 0.074, R^2_c = 0.726$. The test shows that the participants tended to fixate their gaze more when performing sub-tasks looking at the overlapping areas ($\bar{x}_{Overlap} = 0.198m^2, sd_{Overlap} = 0.151m^2$) than when not ($\bar{x}_{\sim Overlap} = 0.226m^2, sd_{\sim Overlap} = 0.173m^2$).

Touchpad Scroll Trajectory Analysis

We analyzed how the participants scrolled the touchpad to see how the participants used touch gestures to support their scanning behaviours. There are 12 over-scrolled trajectories determined by the video analysis and 25 trajectories where no scrolling occurred ($MSD = 0m^2$). These trials were excluded from the first scroll trajectory analysis (Table 5.3-b). The ANOVA indicated that only Technique is statistically significant. Nakagawa's R^2_{GLMM} (ψ_1) are: $R^2_m = 0.093, R^2_c = 0.324$ which indicates a some impact. We found that the participants tended to move the AR screen more with Polyline ($\bar{x}_{MSD} = 0.438m^2, sd = 0.392m^2$) than Radial ($\bar{x}_{MSD} = 0.350m^2, sd = 0.404m^2$). Fig. 5.11 shows the MSD distributions.

Since Radial induced less scroll movement, we investigate if the technique was responsible for the no-scrolling trajectories. Fig. 5.12 shows the distributions of no-scrolling trajectories. We performed a Type II ANOVA with a logistic regression model with MSDNotZero as the response value. MSDNotZero is true if a trajectory's scroll MSD is more than zero. Over-scroll trajectories were included in this analysis. The effects are: Technique ($\chi^2_1 = 0.044, p = 0.833, OR\ (Odd\ Ratio) = 48.866$), Map ($\chi^2_1 = 1.451, p = 0.228, OR = 210.199$), and the interaction between the two ($\chi^2_1 = 0.874, p < 0.350, OR = 0$). The participants are the random effect. Despite having a high adjusted-R^2_{LR} of 0.610 (See Magee [1990] for more information on R^2_{LR}), the result of the test is not significant.

To determine if Radial's overlap feature affected trajectories, we performed a

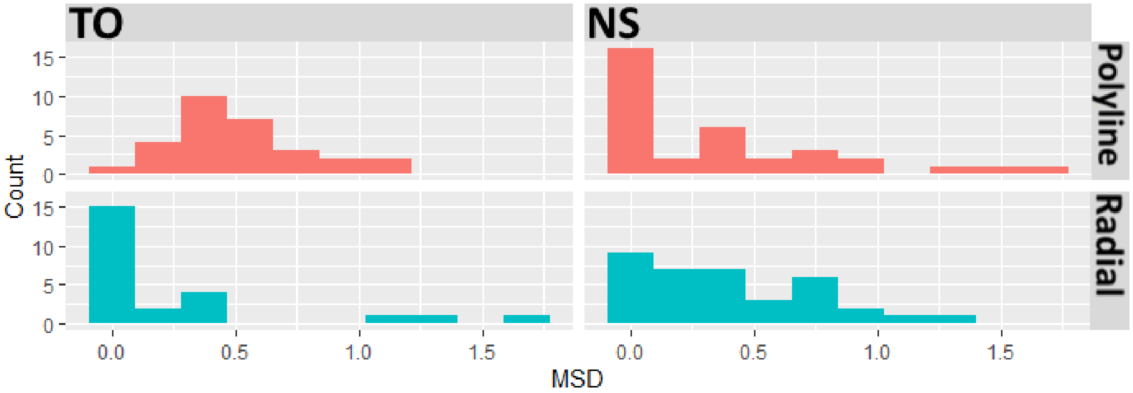


Figure 5.11: The histograms representing the distribution of touchpad scroll MSDs. The unit is square metre (m^2). Bin size = 10.

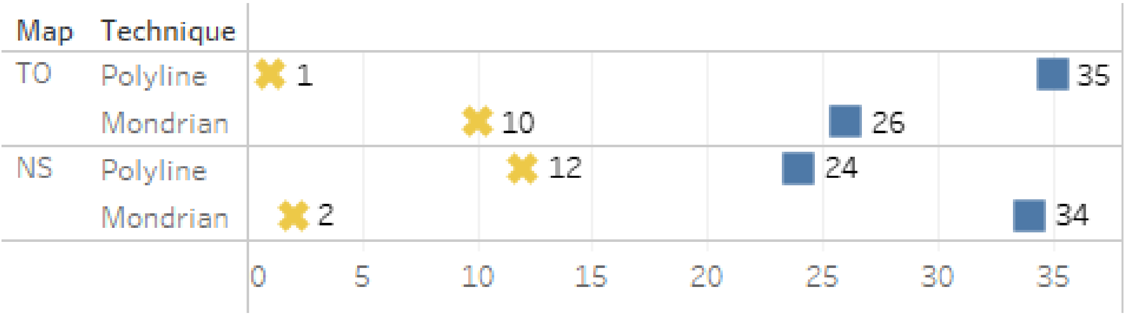


Figure 5.12: The frequencies of scroll v. no-scroll. X denotes that the participants did not scroll. Square denotes that the participants did. Over-scrolled trajectories are included.

Type II ANOVA (Wald's test) with a mixed-effect log-linked Γ GLM with MSD as the response value. Only trajectories that met the following criteria were used: (1) no over-panning, and (2) scroll $MSD > 0$. The fixed effects are: Overlap ($\chi^2_1 = 0.002, p = 0.967, \eta^2 < 0.001$), and Map ($\chi^2_1 = 0.015, p = 0.902, \eta^2 = 0.001$), and their interaction ($\chi^2_1 = 0.046, p = 0.830, \eta^2 = 0.001$). The participants were the random effect. Nakagawa's R^2_{GLMM} (ψ_1) are: $R^2_m = 0.002, R^2_c = 0.675$. The participants did not typically rely on panning to better view the overlapping areas.

Position Trajectory Analysis

We computed MSDs of the HoloLen's 3D positions without any data trimming. The ANOVA (Table 5.3-c) shows an antagonistic interaction effect between Technique and

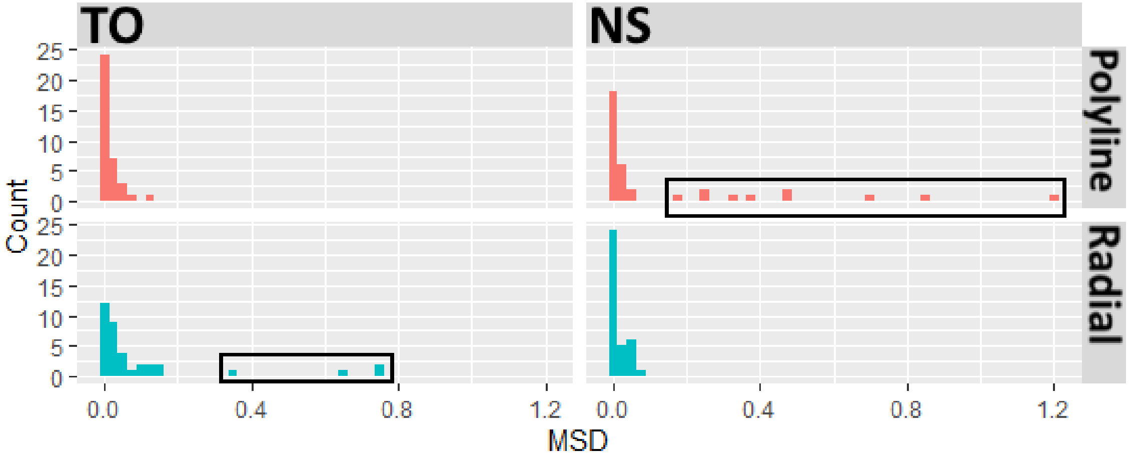


Figure 5.13: Histograms presenting the MSD for the Technique and the Trial. The unit is square metre (m^2). Bin size = 0.025. Extreme MSDs are highlighted in black boxes.

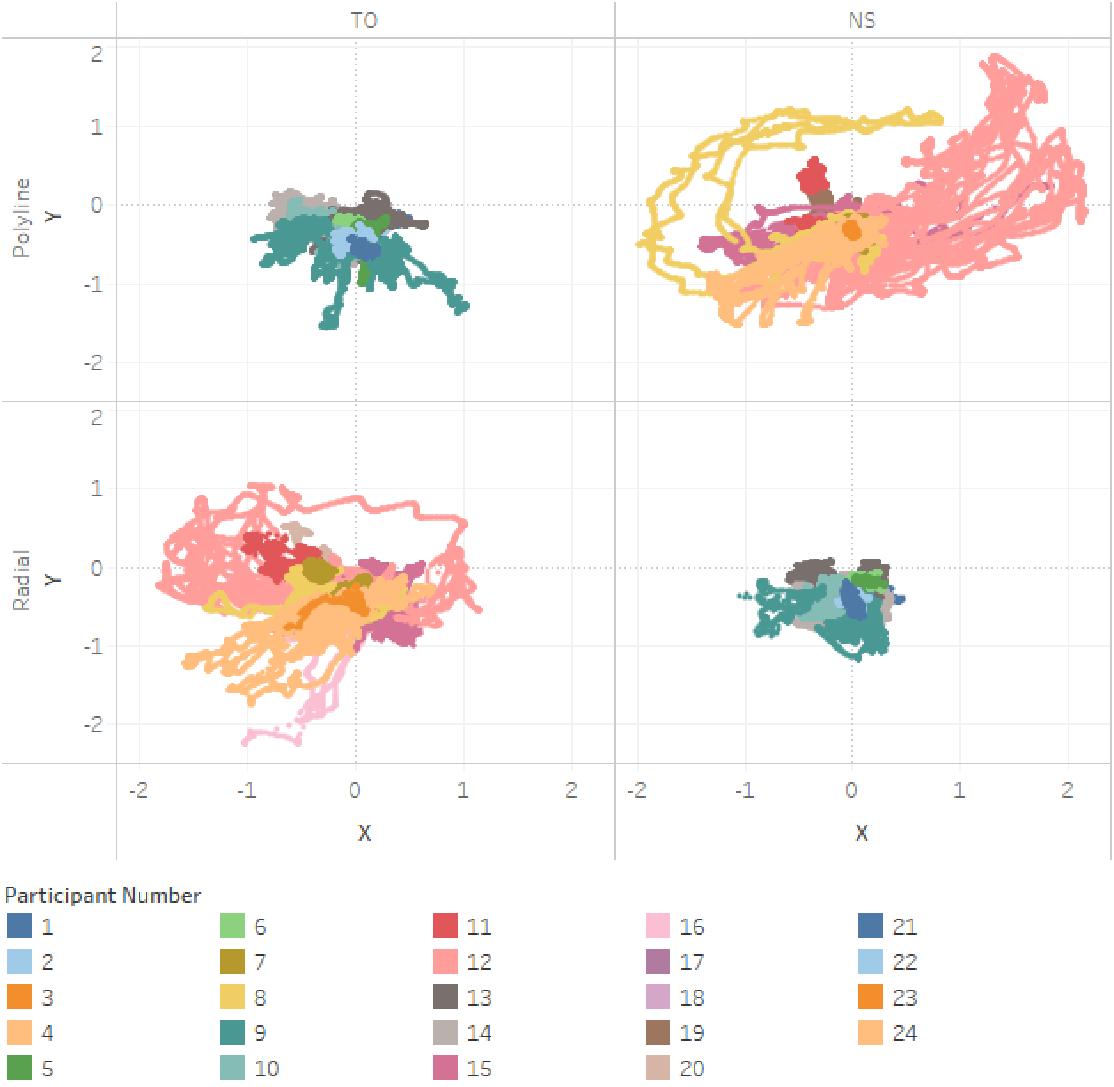


Figure 5.14: Combined trajectories of the HoloLens' X and Y positions grouped by Map, Technique, and participant number. X and Y's units are metre.

Map. Nakagawa's R^2_{GLMM} (ψ_1) are: $R^2_m = 0.165, R^2_c = 0.654$. Based on Fig. 5.13, we made a conjecture that some participants may have moved more than others. To confirm the conjecture, we created and examined Fig. 5.14. We found that particular participants tended to move more than the others during the study. Due to the partially between-subject design of the study, the participants did not perform the tasks with TO+Polyline, and NS+Radial. Therefore, the MSDs of these conditions appeared smaller. The descriptive statistics were: $\bar{x}_{TO+Polyline} = 0.016m^2$ $(sd = 0.024m^2)$, $\bar{x}_{NS+Polyline} = 0.150m^2$ $(sd = 0.278m^2)$, $\bar{x}_{TO+Radial} = 0.105m^2$ $(sd =$

$0.198m^2$), $\bar{x}_{NS+Radial} = 0.016m^2$ ($sd = 0.019m^2$).

Trial Duration Analysis

We performed an ANOVA (Table. 5.3-d) similar to the previous trajectory analyses with an extra factor: Post-fit. Post-fit indicates if the trajectory was pre- or post-fit. The other trajectory ANOVA models do not contain this due to fitting errors. Only Post-fit played a significant role in the trial duration with Nakagawa's R_{GLMM} (ψ_1) effect sizes being: $R^2_m = 0.350, R^2_c = 0.491$. The mean duration for pre-fit tasks was 290.526 seconds with $sd = 99.380$ seconds. The mean duration for post-fit trials was 173.932 seconds with $sd = 76.892$ seconds. This result was indicative of an ordering effect based on tasks as other effects were not statistically significant.

5.3.3 Interview Analysis

To answer RQ2, we analyzed the interview data using the 'bottom-up' approach [Braun and Clarke, 2006]. Overall, the pre-fit and post-fit tasks were easy to perform (n=4) albeit with an initial steep learning curve (n=5). The expanded AR screen was helpful (n=3). For the techniques, the participants (n=10) thought Mondrian was better for identifying correlation. However, Polyline was easier to understand in general (n=13) and felt more precise to use (n=2). Fifteen participants indicated that they experienced colour issues with Mondrian. Some participants (n=4) found the divergent colourmaps difficult to understand and suggested changing them. For instance, P6 believed that additional hues would have been helpful. He stated "eight different, maybe 10 [hues]" could be ideal. P12 thought the colourmaps should be grayscale instead.

5.3.4 Self-Reported Measures

NASA-TLX and SUS

To answer RQ2, we administered NASA-TLX to measure cognitive load. A mixed-effect model parametric ANOVA for NASA-TLX with Technique, Map, Task Types (2-Var, 3-Var, and Postfit) as the fixed effect and the participants as the random effect; a residual analysis with a QQ-plot determined this type of ANOVA was appropriate. Table 5.4 describes the results. Only Technique was statistically significant. The descriptive statistics for Techniques were: $\bar{x}_{Polyline} = 48.793, sd_{Polyline} = 21.472, \bar{x}_{Mondrian} = 52.764, sd_{Mondrian} = 21.582$. Contrary to our expectation, Radial's cognitive load was slightly higher than Polyline's. The interview data indicated that colour distortion may be the cause.

Effects	**SS**	**MSE**	**df1**	**df2**	**F**	**p**	η^2
T	601.79	601.79	1	109.245	4.739	0.032*	0.039
Map	9.11	9.11	1	109.245	0.072	0.789	0.001
TT	146.96	73.48	2	109.321	0.579	0.562	0.010
TT x Map	99.99	99.99	1	41.267	0.787	0.380	0.007
TT x TT	361.61	180.81	2	109.468	1.424	0.245	0.024
Map x TT	51.85	25.93	2	109.239	0.204	0.816	0.003
T x Map x TT	42.23	21.12	2	109.243	0.166	0.847	0.003

Table 5.4: ANOVA for NASA-TLX questionnaire. T = Technique. TT = Task Type (Pre-fit with two variables v Pre-fit three variables v Post-fit). * denotes $p \leq 0.05$.

We rescaled the SUS score using Lewis & Sauro's method [Lewis and Sauro, 2017] since the 10th SUS question was missed in error. Using ART-ANOVA on SUS with Technique as the fixed effect and the participant as the random effect, we found statistical significance ($\bar{x}_{\Delta} = 7.17, F_{1,23} = 6.575, t_{23} = 2.564, p = 0.017, d = 0.741$). We found the median SUS scores are 72.22 for Polyline and 63.888 for Radial.

Effect Size Questionnaires

To answer RQ3 and to understand how well the participants comprehended MLR information conveyed by the glyphs, we administered ESQs after the participants had completed the tasks. The questions of the ESQ varied based on the type of task (pre-fit v. post-fit) and extended the ones found in Peña-Araya et al. [2020]. The questions were:

Pre-fit Q1 (4 levels) Is there any correlation in the data?

Pre-fit Q2 (4 levels) Is there any spatial autocorrelation in the data?

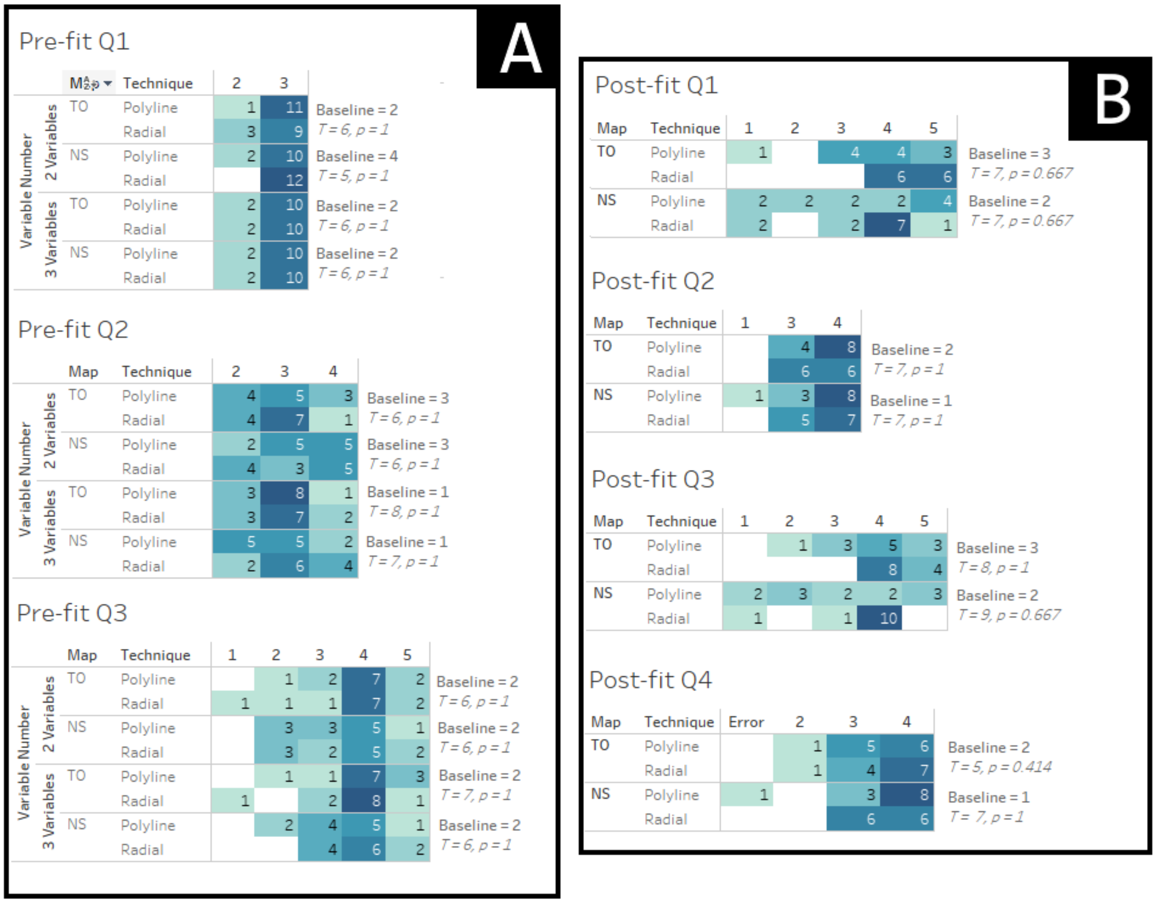
Pre-fit Q3 (5 levels) Is there any multiplicative effect in the data?

Post-fit Q1 (5 levels) What is the goodness-of-fit for Model 1?

Post-fit Q2 (4 levels) Is there any spatial autocorrelation for Model 1's goodness-of-fit?

Post-fit Q3 (5 levels) What is the goodness-of-fit for Model 2?

Post-fit Q4 (4 levels) Is there any spatial autocorrelation for Model 2's goodness-of-fit?



Pre-fit Q1

Variable Number	Map	Technique	2	3	
2 Variables	TO	Polyline	1	11	Baseline = 2
		Radial	3	9	*T = 6, p = 1*
	NS	Polyline	2	10	Baseline = 4
		Radial		12	*T = 5, p = 1*
3 Variables	TO	Polyline	2	10	Baseline = 2
		Radial	2	10	*T = 6, p = 1*
	NS	Polyline	2	10	Baseline = 2
		Radial	2	10	*T = 6, p = 1*

Pre-fit Q2

Variable Number	Map	Technique	2	3	4	
2 Variables	TO	Polyline	4	5	3	Baseline = 3
		Radial	4	7	1	*T = 6, p = 1*
	NS	Polyline	2	5	5	Baseline = 3
		Radial	4	3	5	*T = 6, p = 1*
3 Variables	TO	Polyline	3	8	1	Baseline = 1
		Radial	3	7	2	*T = 8, p = 1*
	NS	Polyline	5	5	2	Baseline = 1
		Radial	2	6	4	*T = 7, p = 1*

Pre-fit Q3

Variable Number	Map	Technique	1	2	3	4	5	
2 Variables	TO	Polyline		1	2	7	2	Baseline = 2
		Radial	1	1	1	7	2	*T = 6, p = 1*
	NS	Polyline		3	3	5	1	Baseline = 2
		Radial		3	2	5	2	*T = 6, p = 1*
3 Variables	TO	Polyline		1	1	7	3	Baseline = 2
		Radial	1		2	8	1	*T = 7, p = 1*
	NS	Polyline		2	4	5	1	Baseline = 2
		Radial			4	6	2	*T = 6, p = 1*

Post-fit Q1

Map	Technique	1	2	3	4	5	
TO	Polyline	1		4	4	3	Baseline = 3
	Radial				6	6	*T = 7, p = 0.667*
NS	Polyline	2	2	2	2	4	Baseline = 2
	Radial	2		2	7	1	*T = 7, p = 0.667*

Post-fit Q2

Map	Technique	1	3	4	
TO	Polyline		4	8	Baseline = 2
	Radial		6	6	*T = 7, p = 1*
NS	Polyline	1	3	8	Baseline = 1
	Radial		5	7	*T = 7, p = 1*

Post-fit Q3

Map	Technique	1	2	3	4	5	
TO	Polyline		1	3	5	3	Baseline = 3
	Radial				8	4	*T = 8, p = 1*
NS	Polyline	2	3	2	2	3	Baseline = 2
	Radial	1		1	10		*T = 9, p = 0.667*

Post-fit Q4

Map	Technique	Error	2	3	4	
TO	Polyline		1	5	6	Baseline = 2
	Radial		1	4	7	*T = 5, p = 0.414*
NS	Polyline	1		3	8	Baseline = 1
	Radial			6	6	*T = 7, p = 1*

Figure 5.15: The frequency tables of the scores for the pre-fit ESQs **A:**, and post-fit ESQs **B:**. 1 = weakest, 5 = strongest. Baseline numbers and the permutation test statistics for comparing the techniques are provided on the right margin of each table. Some scores with the frequency of zero in all rows have been removed. Error means the participant did not answer.

For the 4-level questions, the levels were: (1) none, (2) weak, (3) moderate, and (4) strong. These levels are based on Cohen's interpretation of effect size [Rosenthal, 1996]. For the 5-level questions, the levels were: (1) very weak, (2) weak, (3) medium, (4) strong, and (5) very strong. These questions were for novel effect sizes. Fig. 5.15 contains the distributions of the participants' answers. The participants tended to overestimate their answers for both techniques. None of the permutation tests for comparing the techniques is statistically significant–meaning that glyph-based visualization in general may yield overestimated results regardless of visual channels used, and pre-existing colour display issues inherent to OST-HWDs. Furthermore, Radial's overlapping areas did not seem to improve the participants' estimation of the multiplicative effects.

To create the baseline values in Fig. 5.15, we computed various statistics. For Pre-fit Q1, we computed the mean R^2 for the variables:

- TO (2 Variables): 0.157
- TO (3 Variables): 0.269
- NS (2 Variables): 0.809
- NS (3 Variables): 0.221

For Pre-fit Q2, we computed Moran's I which are available in Table. 5.5. For Post-fit Q1, Q3, we created mean estimates are based on mean E_L:

- TO Model 1: 0.404
- TO Model 2: 0.407
- NS Model 1: 0.281
- NS Model 2: 0.286

For Pre-fit Q3, we computed the following values for multiplicative effects:

- TO (2 Variables): 0.563
- TO (3 Variables): 0.453
- NS (2 Variables): 0.090
- NS (3 Variables): 0.088

The Moran's I statistics for Post-fit Q2, and Q4 are as following:

- TO Model 1: $I = 0.088, Expected = -0.004, sd = 0.012, p = 0.000$
- TO Model 2: $I = 0.095, Expected = -0.004, sd = 0.012, p = 0.000$
- NS Model 1: $I = 0.014, Expected = -0.004, sd = 0.012, p = 0.122$
- NS Model 2: $I = 0.032, Expected = -0.004, sd = 0.012, p = 0.003$

We performed permutation tests with `exactRankTests` to see if there is any difference between the techniques as recommended by Collingridge [2013]. Fig. 5.15 shows the permutation test statistics on the top-left corners of each associated chart. None

			Moran's I Information			
Variables	**Map**	**Layers (Pre-fit)**	***I***	***Expected (Null)***	***sd***	***p***
Security	Apartment	2 Var	0.152	-0.003	0.017	0.000
Stairwells	Apartment	2 Var	0.067	-0.003	0.017	0.000
Graffiti	Apartment	3 Var	0.076	-0.003	0.017	0.000
Exterior Cladding	Apartment	3 Var	0.075	-0.003	0.017	0.000
Exterior Ground	Apartment	3 Var	0.121	-0.003	0.017	0.000
Calcium	Lake	2 Var	0.158	-0.004	0.013	0.000
Chloride	Lake	2 Var	0.137	-0.004	0.013	0.000
Iron	Lake	3 Var	0.003	-0.004	0.013	0.010
Manganese	Lake	3 Var	0.153	-0.004	0.013	0.000
Potassium	Lake	3 Var	0.088	-0.004	0.013	0.000

Table 5.5: Moran's I statistics for the variables used in the pre-fit stage.

of the tests is statistically significant. We find no evidence that Technique affected participant answers. Fig. 5.15 further shows that the participants tended to overestimate their answers regardless of the techniques. This is suggestive of a universal flaw in glyph-based visualization which warrants additional studies.

5.4 Limitations

Our work has two major limitations. First, we only considered 2D visualization. While Radial glyphs could be composed in other directions (e.g., vertical) to support 3D visualization, we limited them to 2D arrangement so Radial could be compared to Polyline, a strictly 2D technique. Secondly, while our tasks provide a good balance between experimental control and external validity, we should supplement it with an additional study like the one by Jankun-Kelly et al. [2010]. In their study, the participants only analyzed one glyph per trial. An expanded version of the interface should be qualitatively evaluated by experts who practice geospatial analyses.

5.5 Discussion

We identify three main discussion topics: (1) scanning behaviour, (2) the relationship between the AR and the tablet, and (3) colour and usability. The first two topics touch on RQ1 as they pertain to how the techniques affected scanning behaviours. The second topic focuses on how HoloLens's display technologies affected the usability of Radial, answering RQ2. Unfortunately, we cannot answer RQ3.

5.5.1 Scanning Behaviour

We found Polyline and Radial to induce different scanning behaviour. The former tends to induce more scrolling on the tablet, while the latter induces more gaze scanning. Therefore, in general, we should use a shape-based technique to encourage closer examination of glyphs, while colour-based techniques are better for encouraging cursory explorations at the pre-attentive level. However, tasks can also affect gaze fixation; we found the participants' gazes tended to become more fixated on Radial glyphs when looking at the overlapping areas. Since the test of ESQ did not reveal any difference between techniques, we cannot answer RQ3 and state if either technique is superior in terms of accuracy. However, the ESQ results are still interesting as they show the participants tended to overestimate values with both techniques–hinting at another research direction: can glyphs encourage overestimation, and if so, how?

5.5.2 Tablet: A Display Device and/or an Input Device?

In general, the participants tended not to rely on the tablet for display–despite the tablet providing more visual information. This suggests that the tablet should be more of an input hardware, and used similarly to the mobile devices in STREAM [Hubenschmid et al., 2021]. This also shows that having a focus+context (F+C) hardware arrangement does not necessarily guarantee a F+C interface. F+C is a type of display paradigm with two display resolutions: one high resolution for where the user is focusing, and one low resolution for contextual information around the focus area [Baudisch et al., 2001]. With the tablet having a superior display quality, it follows that our interface should be F+C. However, the participants tended not to focus on the tablet–rendering the paradigm moot. Neither Polyline nor Radial had any effect on how the participants used the AR and the tablet together. In the future, we may need to explore other display paradigms instead like overview+detail (O+D), where the tablet displays overview information (e.g., mini-map), and the AR displays more details [Yang et al., 2022]. In the chapter describing the redesign of Gander (Ch 9), we a more flexible paradigm which allows the user to switch between O+D and F+C. The user, when interacting with the visualization from afar, uses O+D to select and segment virtual content. Meanwhile, the user can also place the tablet onto the virtual objects to activate the F+C mode in order to glean more information from the objects.

5.5.3 Colour and Usability

The study shows that the HoloLens colour display is problematic enough to affect the usability of Radial, which may explain the lower NASA-TLX and SUS scores. Some participants, like P6, stated that a multi-hue colourmap, like the rainbow colourmap, might have made the tasks easier. While this contradicts guidelines [Crameri et al., 2020], studies [Quinan et al., 2017, Reda and Szafir, 2021] suggest that users actually tend to mentally dissect a colourmap into smaller segments. A rainbow colourmap could facilitate value assessment by making it easier for the user to dissect the colourmap. The colourmap also may be more effective with an additive OST-HWD like the HoloLens.

5.5.4 New Effect Size Questionnaire

The tests on ESQ do not show any difference between the techniques. While increasing the sample size may help to achieve statistical significance, we must rethink how we collect the data. Looking at the video data, we found the participants tended to focus on local areas of the maps. This means the participants may have provided answers based on local areas, and not global information. A better way to collect effect size data is to administer an in situ AR-based adaptive questionnaire. The questions should be local to the participants' work area. The administration should also happen while the participants are completing the tasks. A similar example of this is the work by Yang et al. [2021] where participants provided the answer in AR. We must also reconsider the levels in the questions; four to five levels are unlikely to be sufficient to distinguish two techniques. However, we cannot simply ask participants to provide the exact number (e.g. "Please indicate R^2 as a number between 0 and 1"); our pilot study indicated that this type of question would confuse the participants. We could also consider the use of artificial intelligence (AI) to seamlessly monitor and collect results from participants.

5.6 Conclusion

This chapter describes the synoptic study which compared two glyph visualization techniques using semi-naturalistic tasks and conditions. For Obj1, the study results demonstrate that the tasks themselves did not have much effect on the trajectory data. This is somewhat expected, because we designed the pre-fit tasks to be similar to the post-fit tasks. For Obj2, unfortunately, the ESQ was not sensitive enough to measure the effectiveness of the techniques. To address this issue, we must develop a new data collection method. Our results are the most fruitful for Obj3, we found that different visualization techniques can influence how the user scans the glyph fields. The results also demonstrate that having two display resolutions is not enough to have a F+C. We must also enforce user behaviour.

To better understand the nature of pre-fit and post-fit tasks, and to better fulfill Obj1, we conducted the walkthrough demonstration study (Chapter 7). The study participants were experts who would use the study. We conducted the elementary study (Chapter 6) with a much higher degree of control, and a much simpler task so that we can better understand how the user can glean information from a single composite. The results of this study may inform how to design future versions of the ESQ.

Chapter 6

The Elementary Study: Comprehension of Individual Glyph Composites

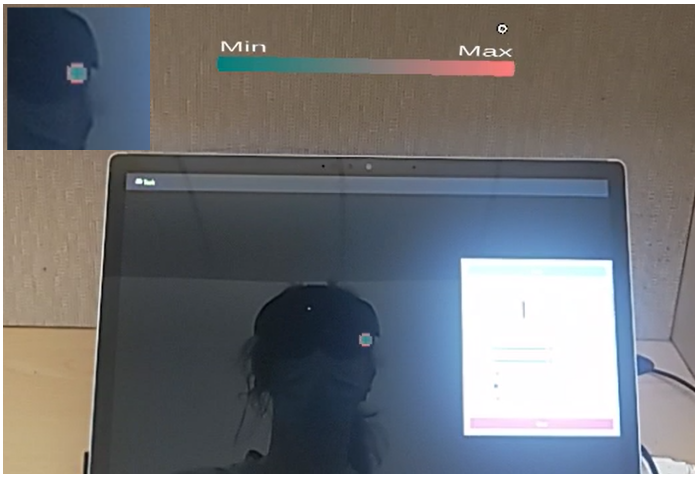


Figure 6.1: P6 completing a trial. Radial glyph composite with four constituent glyphs. The top-left inset shows the zoomed-in version of the glyph.

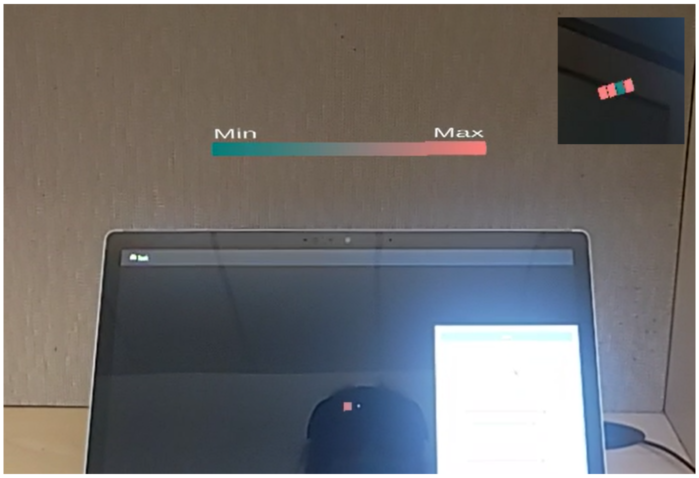


Figure 6.2: P6 completing a trial. Radial glyph composite with four constituent glyphs. The top-right inset shows the zoomed-in version of the glyph.

The purpose of the elementary study is to fill one of the gaps identified in the synoptic study. The synoptic study's instrument is insufficiently sensitive for us to understand how each participant obtains information from individual glyphs. Therefore, we conducted this elementary study with a much simpler and more controlled task. Our study design is based on the work of Jankun-Kelly et al. [2010] where each participant indicates specific values found in a glyph. In that study, each trial only contained one glyph–making the study design elementary based on Andrienko and Andrienko [2005]'s task classification. Hence, we call this study "the elementary study." Due to the study's focus on glyph comprehension, this study primarily fulfills Obj2. Furthermore, the synoptic study did not compare Stacked (Fig. 6.2, and described in Ch. 4). In order to avoid having the parallax effect as a confounding variable, the synoptic study excluded Stacked.

We conducted a study exploring the perception of the parallax effect on glyphs on an immersive augmented reality (AR) display. Unlike existing work on comprehending glyph fields [Fuchs et al., 2017], we focus on the comprehension of a single glyph composite at a time. While studies exploring glyph fields help build an understanding of how a user may extract higher-level information (such as trends or correlations), their results do not examine lower-level operations – in particular, how the user perceives and derives information from individual glyphs or glyph composites. Additionally, head-worn AR displays present new challenges and opportunities for glyph comprehension. First, they facilitate 3D visualization techniques, including 3D approaches to arranging glyphs. Second, they permit very large areas for visualizations, which can engender more direct forms of physical movement in relation to data and raise considerations regarding the perception of visual information at a distance. These challenges and opportunities require careful examination. Evaluations of complete AR visualization systems in the lab or in the field tend to produce results that are difficult to generalize [McGrath, 1995]. The instruments used in such studies are typically not sensitive enough to examine fine-grained actions involved in visual comprehension. We also require controlled studies that focus on visual comprehension and basic interaction using elementary tasks. According to Andrienko and Andrienko [2005], elementary tasks in information visualization are ones that only require an understanding of local information.

Glyph-based visualization is a visualization technique that represents spatial data using visual markers. For instance, we can use glyphs to represent vectors in a field [Rocha et al., 2017] or points of interest on a map [Peña-Araya et al., 2020]. In any visualization, there are multiple visual properties (or visual channels) that we can alter to express different values [Borgo et al., 2013]. Our channels of interest in this study are positions, intersection, and colour. We use position to represent different distances of the glyphs from the user. Since immersive analytics often makes use of augmented reality (AR) devices, a glyph can be close to the user or very far from the user, impacting its visibility. By intersection, we mean whether and how glyphs overlap. In this study, we explore two types of intersections: Radial, and Stacked. Colour is a dominant visual channel [Borgo et al., 2013] which makes it more immediately noticeable than some other channels. According to Fuchs et al. [2017], there are very few empirical glyph-based studies that solely focus on colour.

According to Ropinski et al. [2011], glyph composition is a combination of multiple glyphs, used to convey multivariate data. We can use a composite to avoid overdrawing. A well-designed glyph composition will effectively arrange multiple glyphs, each representing a different variable or attribute, around a single position. In our study, we compare two glyph composition techniques. Radial, a 2D composition technique, arranges the constituent colour-based glyphs in a partially overlapping radial pattern. The colour at the center of the radial composite is determined by composing the values of individual glyphs (e.g., by multiplying or averaging their terms, applying additive blending, showing the maximum value, etc.). Stacked, a parallax-based composition technique, arranges the glyphs by vertically stacking them on top of each other. Stacked glyph composites are spatially separated from each other. This allows one to compose or decompose the composite through the parallax effect (see Rouan [2015]) by changing their viewing angles. We designed both Radial and Stacked to be used in multivariate visualizations using an OST-HWD to create a larger glyph field. Design considerations included balancing glyph complexity with processing power, and determining colour scales that can be rendered effectively on an OST-HWD.

We conducted a repeated-measures experiment with 16 participants. In each trial, each participant indicated the colour values of the glyphs in a composite. We compare accuracy, efficiency, and perceived usability when acquiring information from Radial vs. Stacked glyph composites. We define accuracy as the absolute difference (AbsDiff) between the true value presented by a glyph, and the participant's selected value, as do Yang et al. [2021]. We measure efficiency as the time spent identifying the values of all glyphs in a single composite. We measured perceived usability using UMUX (see Lewis et al. [2013]) to gauge how the techniques affect how the participants feel about them. Between trials, we varied the number of glyphs in a composite and the distance of the glyph composite from the participant.

Our study design is similar to that of a study conducted by Jankun-Kelly et al. [2010]. Their study also involved elementary value-judgment tasks which allowed them to understand how individual visual attributes or channels contribute to the understanding of the glyph. There is a small difference, however. While their study presents a single glyph per trial and the participants indicated the values of different visual channels, our participants indicated the colour values of multiple glyphs per trial. Nevertheless, our glyphs were composed into a single unit for analysis.

We found the glyph composition sizes (GCS) to play the most important role. As GCS increases, the composites become harder to understand, making both techniques less accurate. GCS had more effect on Stacked than Radial. We suspect the participants' inherent bias to read glyphs in reverse order may be behind this. More work is necessary to verify this bias. When glyphs were close to the participants, neither Radial nor Stacked were faster than each other, despite Stacked eliciting more head movements. This means the ability to decompose glyph composites may have a compensatory effect. When the glyphs were further away, Stacked was faster than Radial, because the glyphs were already decomposed. In terms of usability, both techniques were not different. The participants felt they moved slightly more with Stacked. Although our study was not mainly about colourmap design, the results here could be used for additional research into the subject.

This work is possible thanks to advancements in display technologies. Without an untethered, head-worn, AR display, designing novel visualization techniques that use the parallax effects and studying them would have been impractical. Our glyph design work serves as a foundation for incorporating the parallax effect and the blending effect into immersive analytics.

6.1 Glyph Composition Techniques

In this section, we described two types of glyph composition techniques in the study. Fig. 6.3 provides a 2D diagram on how glyphs are composed based on the techniques.

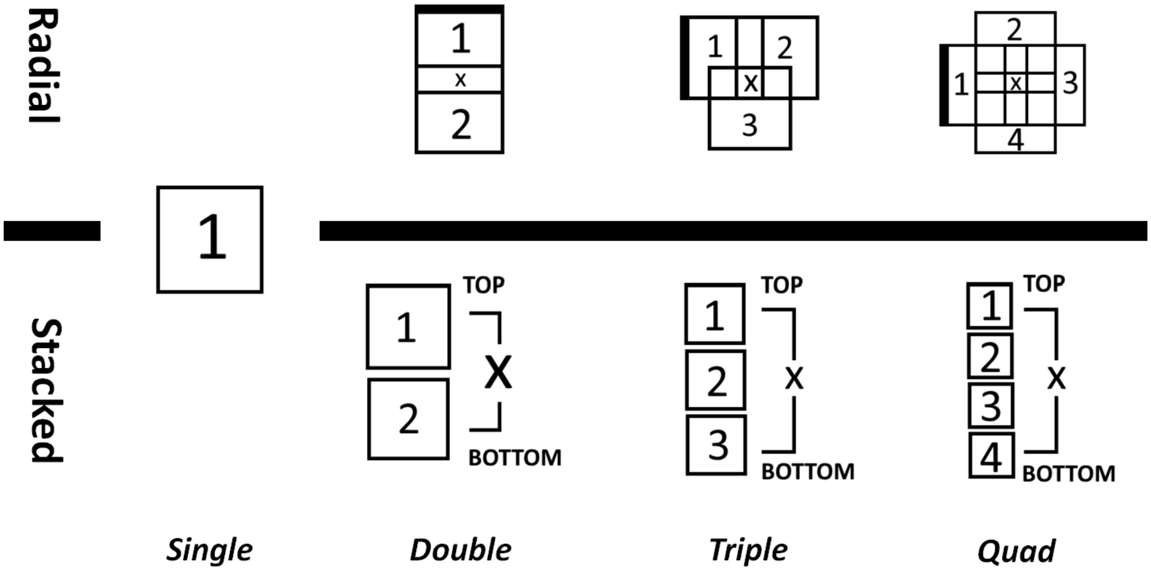


Figure 6.3: The designs of Radial and Stacked with respect to GCS. Radial and Stacked are indistinguishable when GSC=1 (Single). X denotes the area where the user can find the blended value of all glyphs; for Stacked, the blended value requires the user to change their viewing angle to achieve a parallax effect. While some diagrams of Radial show thicker borders on some glyphs, the actual glyphs do not have any border. Since we also use these diagrams in the study software interface, we added thickness to help the participants find the first glyph in the Radial composites.

6.1.1 Radial

Radial (Fig. 6.1) arranges the glyphs in a radial pattern. Each glyph is offset from the centre of the composite. The amount of offset is 360^o divided by GCS. After obtaining the angles, we create offset direction vectors for the glyphs which must be on the same 3D plane. Each vector is 0.15cm (1.5mm) long. For our study, we define the horizontal plane as the surface of the tablet's screen. The overlapping areas represent the blending of the constituent glyph values. In composites with higher GCS, we can observe more overlapping. The centre overlapping area represents the blending of all glyphs. For more information on arrangement based on GCS, please refer to Fig. 6.3.

6.1.2 Stacked

Stacked (Fig. 6.2) arranges the glyph in a vertical line perpendicular to a plane. Each glyph floats on top of each other at 0.5cm (5mm) apart. While the glyphs are spatially apart, the user can compose the glyphs using changing their viewing angle. Changing one's viewing angle can affect the apparent distance and sizes of objects through the parallax effect, which can introduce occlusion or overlapping between the glyphs. Where the occlusion occurs, Stacked blends the area. If the user wishes to see the blending of all glyphs in a composite, they must visually align all glyphs. When the glyphs are far away, the parallax effect automatically separates the glyphs and the user can no longer view the overlapping areas. Fig. 6.3 provides additional information on the appearance based on GCS.

6.2 Method

We conducted the study with 16 participants (8 Females, 7 Males, 1 Other) recruited from Dalhousie University's daily e-mail communication, and its school of computer science's mailing list. Degrees completed were: high school–*5*, undergraduate–*8*, and graduate–*3*. Nine participants had used OST-HWDs before the study. Only three had participated in studies that used them. We assessed the participants' effectiveness and satisfaction with the techniques. The study conditions were: techniques (Stacked v. Radial), values, GCS, and distances from the participants. There were four categories of hypotheses: accuracy, time, and usability and body movement.

The hypotheses for accuracy (H1.1-H1.4) are:

- **H1.1: Stacked yields more accurate value estimations than Radial.** Radial presents glyphs in a partially overlapping fashion on the same horizontal plane. According to Moreland [33], looking at neighbouring colours at once can skew the perception of the colours.

- **H1.2: Increasing glyph composition size (GCS) does not affect accuracy for the Stacked technique.** Because individual glyphs are spatially separable with the Stacked technique, increasing GCS should not impact colour value estimation accuracy.

- **H1.3: Increasing glyph composition size (GCS) decreases accuracy for the Radial technique.** Increasing GCS increases the number of adjacent colours, which could impact colour value perception.

- **H1.4: Increasing the distance between a glyph and the viewer reduces accuracy, for both techniques.** When glyphs are further away, they appear smaller to the viewer, and are viewed at a more oblique angle. These factors will negatively impact colour estimation accuracy. In this study, we have three levels of distances: 0m, 0.2m, and 1.5m. 0m represents the case where a composite is readily visible to the user. 0.2m represents the scenario where there is ambiguity if a composite is mounted relative to the user or to the environment. A far-away composite is represented by the 1.5m distance.

The hypotheses for trial duration (H2.1-H2.3) are:

- **H2.1: Radial will be more efficient than Stacked for deriving glyph colour values near the viewer.** Stacked glyphs require head movement to manually decompose the glyph, increasing the time required to obtain individual glyph colour values.
- **H2.2: Stacked will be more efficient than Radial for deriving glyph colour values further from the viewer.** Increasing GCS also increases the number of glyphs, which effectively lengthens a trial.
- **H2.3: Increasing the distance between a glyph and the viewer increases the time required, for both techniques.** Because colour values are more difficult to discern at a distance, more time will be spent assessing colour values.

The hypotheses for usability and movement (H3.1-H3.2) are:

H3.1: Stacked will be deemed as more usable. Since the user can compose and decompose Stacked glyphs, they find it easier to complete the tasks with the technique.

H3.2: Participants feel Stacked induces more movement. The participants feel that they need to move more to compose or decompose Stacked glyphs.

6.2.1 Apparatus

Each participant was seated in a swivel chair close to a tablet (Microsoft Book Surface 3 with a 15in screen). While the participant would not be able to move the chair, they could rotate the chair to better view the glyphs. They could also move their torsos, and their heads to change their viewing angle. We propped the tablet up so that it would be angled at 60^o. This would allow Stacked glyphs to blend automatically with minimal head rotation from the participant. Since our pilot study indicated that frequent finger swiping could be uncomfortable, the participant used a mouse to answer the questions for the trials. The participant used Microsoft HoloLens v2 to view the glyph composites in AR. Fig. 6.4 shows the physical set-up of the study.

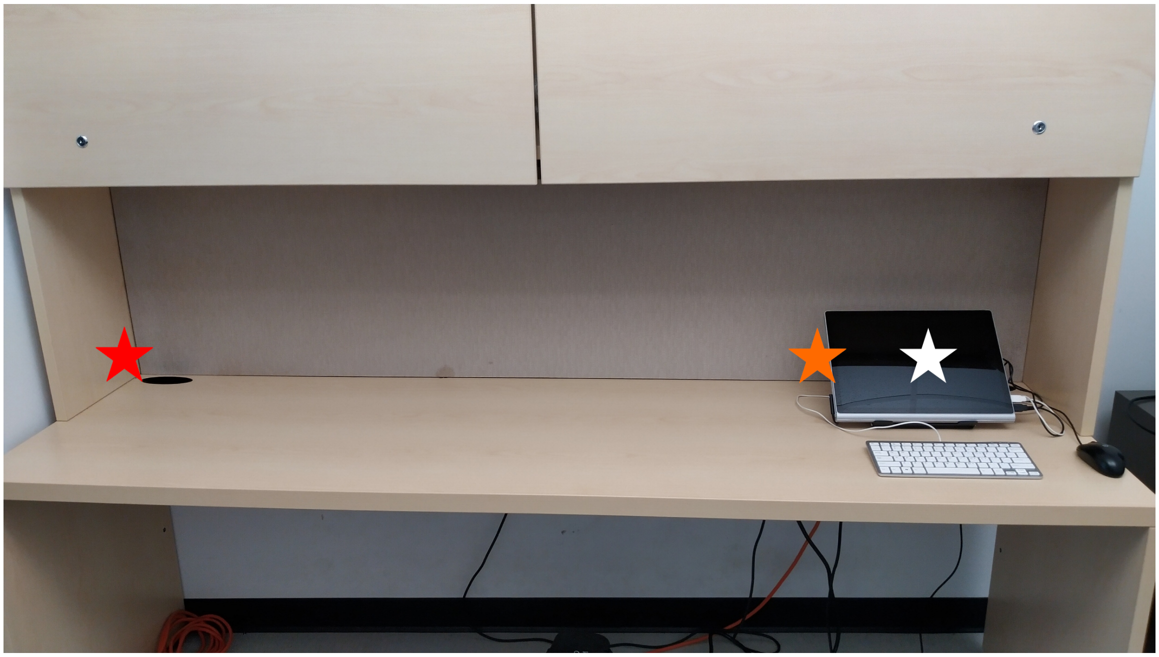


Figure 6.4: The setup of the laboratory for the study. Each star represents the distance of the glyph composites from the centre of the tablet. White star = 0m. Orange star = 0.2m. Red star = 1.5m.

Tablet Interface

The participants indicated the values of the glyphs using the tablet interface, implemented using HTML as seen in Fig. 6.5. The interface consists of a diagram of the glyph design, and the sliders. Fig. 6.3 contains all possible diagrams that the participants will see throughout the study. The participants move the slider to indicate the desired values. Each slider has the minimum value of 0, the maximum value of 1, and the step value of 0.01. We purposely kept the background of the tablet black to minimize the interaction between the light from the tablet's screen and the glyphs. Since HoloLens relies on an additive display, it must modify the light from

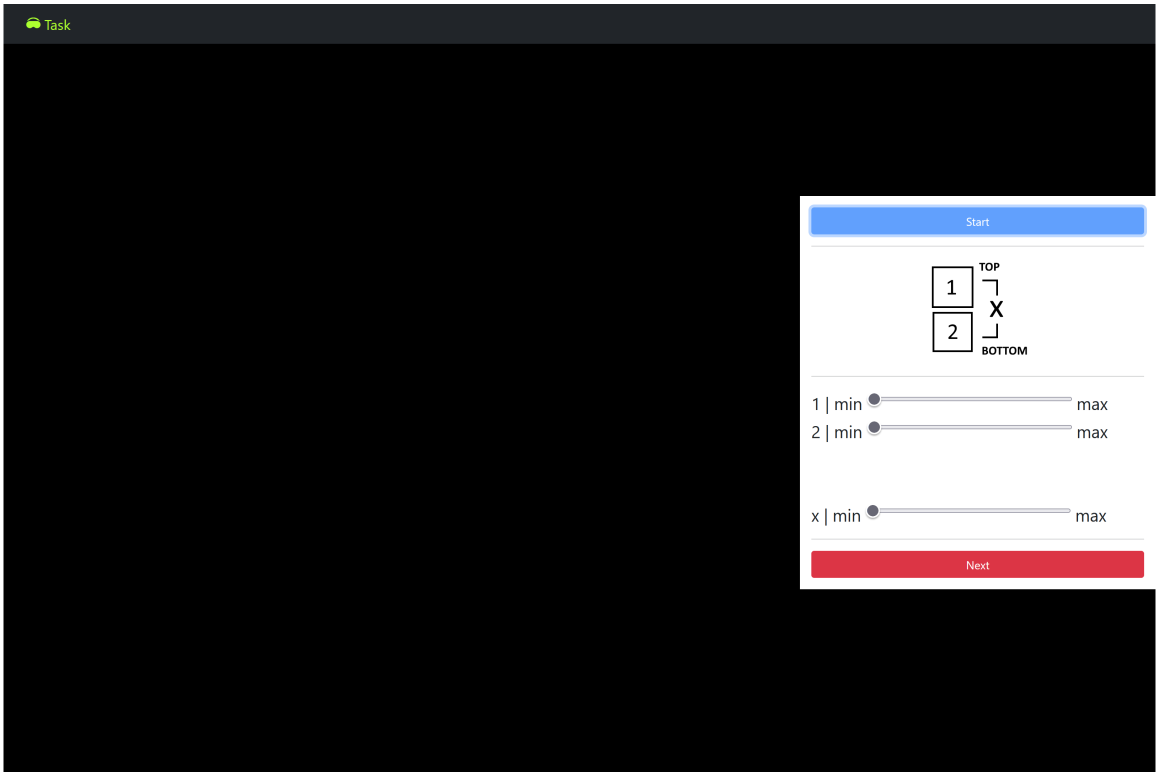


Figure 6.5: The interface of the study software on the tablet.

the physical world into ideal display colours. This is harder when the light from the physical world is very bright. If the glyphs appear off-screen (i,e. at 0.2m, and 1.5m distances), the tablet interface also displays a left arrow alongside the diagram. The tablet's background is predominantly black to prevent the HoloLens v2 from being affected by the tablet's light.

Augmented Reality Interface and Colourmap

The glyphs are rendered in AR and their positions are anchored based on the initial QR synchronization. The colourmap used a modified version of "Conifer Forest", an ArcGIS colourmap [ArcGIS, 2022]. We called our version Saga. The implementation code of Saga is in Chapter 4. We are aware that this colourmap is not accessible for those with colour-vision deficiency (CVD), because it is more isoluminant [Crameri et al., 2020, Kovesi, 2019]—i.e. lower values and higher values have similar luminance [Kovesi, 2019]. However, isoluminance is important in this study due to OST-HWDs like Microsoft HoloLens v2 treating darkness as the same as transparency [Itoh et al., 2021]. If we use a scientific colourmap that is sortable through luminance alone like

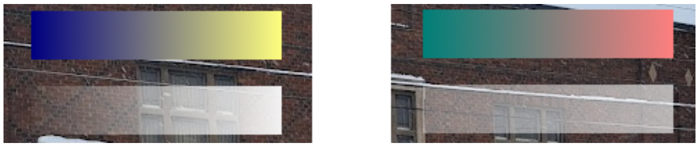


Figure 6.6: LEFT: Ukraine, the default colourmap on the top. Ukraine with its luminance converted to alpha at the bottom. **RIGHT:** Saga, the colourmap used in the study on the top. Saga with its luminance converted to alpha at the bottom. Saga is more isoluminant than Ukraine.

Cividis [Nuñez et al., 2018], the glyphs with lower values would appear less visible to the participants. This behaviour can act as a confounding variable in the study. Fig. 6.6 demonstrates the relative isoluminance between Ukraine and Saga.

6.2.2 Procedure

At the beginning of a session, we administered a 16-plate Ishihara test (see Melamud et al. Melamud et al. [2004]) to each participant to determine if they have colour-vision deficiency (CVD). We warned the participants that the test was only cursory, and not a diagnosis. Everyone passed the test. Then, each participant completed a demographic questionnaire and training trials for the techniques. We assigned the techniques based on the participant's ID number. If they had an odd ID number, they would use Radial first, then Stacked. Otherwise, the order was reversed. During the training trials, GCS progressively increased–starting with Single glyphs, then Double, then Triple, and finally, Quad glyphs. Additionally, we increased the distances from the user–from 0m, to 0.2m, and finally 1.5m left of each participant.

To complete a trial, the participant performed the following:

- **STEP 1.** Clicked on the start button or the words "Click to begin the trial" to show the glyphs.
- **STEP 2.** Visually located the glyph–if the glyph was not at 0m, a left-pointing arrow appeared as a guide on the tablet.
- **STEP 3.** For each constituent glyph, reviewed the corresponding glyph component and indicated its value using sliders presented on the right side of the interface. Additionally, identified the value of the area where all constituent glyphs overlapped.

- **STEP 4.** Clicked on “End Trial.”

STEP 4. Clicked on “End Trial.”

Each participant performed 108 experimental trials with varying: (1) colour values (randomized between 0.0 to 1.0 inclusive, step = 0.1), (2) distances (0m, 0.2m, 1.5m left of the participant), and (3) GCS (Double, Triple, Quad). The order was randomized. Then, they completed UMUX (see Lewis et al. [2013]) which is a set of 7-point Likert scale statements, plus four additional statements. The questionnaire statements were as follows (+ denotes an additional statement.):

Q1 The technique is good overall.

Q2 Using this technique is frustrating.

Q3 The technique is easy to use.

Q4 I spent too much time with the technique.

Q5+ The technique is easy to understand.

Q6+ I moved and rotated my head a lot with the technique.

Q7+ I moved my body a lot with the technique.

Q8+ I prioritized speed over correctness.

Afterwards, the participants trained on a different technique, repeated the 108 trials, and completed the questionnaire again for the new technique. Finally, we paid the participants 15 Canadian dollars for their time.

6.3 Analysis

We performed two omnibus tests: one for accuracy (AbsDiff), and another one for trial duration. We analyzed accuracy by performing a regression analysis with the following model:

- **Dependent Variables:** AbsDiff of Glyph 1 ($d1$), Glyph 2 ($d2$), Glyph 3 ($d3$), Glyph 4 ($d4$), and the multiplication values (dx).

- **Independent Variables:** Technique, GCS, mean-squared displacement (MSD), Distance, two-way interaction (excluding MSD), three-way interaction (excluding MSD).

We computed AbsDiff with this formula: $d = |p - v|, p \in [0, 1], v \in [0, 1]$ where p is a participant's value and v is the true value. We adopted the use of AbsDiff from Yang et al. [2021]. Since AbsDiff is a unit interval number, by default, $d1, d2, d3, d4, dx$ have beta distributions Gupta [2011]. Therefore, according to Anderson [2017], we cannot analyze the values using MANOVA which is very sensitive to violations of assumptions. Instead, we used PERMANOVA Anderson [2017], a nonparametric alternative to MANOVA. To see if head movement could have an extraneous effect in terms of accuracy, we added mean-squared distances (MSDs) of all trials as a factor. We used the MSD formula found in Poupard et al Poupard et al. [2019] and applied it to the HoloLens's positions.

We analyzed the trial durations using ART-ANOVA (see Wobbrock et al. [2011]) and ART-Contrast tests (see Elkin et al. [2021]) as the data were not normally distributed. The variables are:

- **Dependent Variable:** Duration

- **Independent Variables:** Technique, GCS, Distance, and the two- and three-way interactions between the variables.

To test the UMUX and Likert scale results, we used Wilcoxon signed-rank tests as the data were not normally distributed.

6.4 Results

6.4.1 Accuracy: H1.1-H1.4

It is important to note that the accuracy of both techniques is poor overall, regardless of the factors; the overall medians for the AbsDiff for $d1, d2, d3, d4, dx$ are $0.25, 0.22, 0.24, 0.275, 0.09$ respectively. Fig. 6.7 shows that the selected values tended to follow "bathtub" distributions. Fig. 6.8 shows, at first glance, that dx seems to be very accurate–particularly for higher GCS. However, as it turns out, increasing the

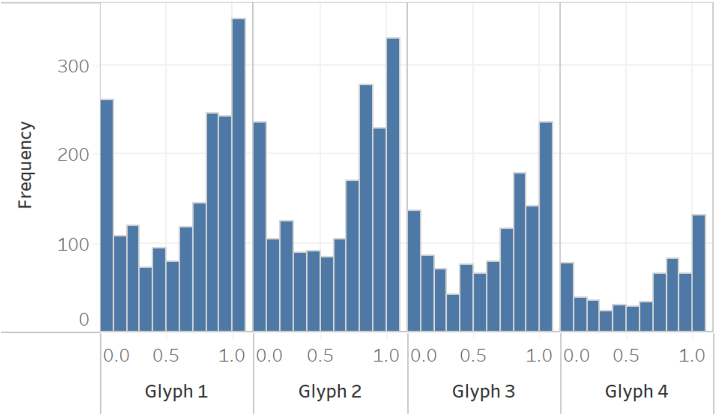


Figure 6.7: The distributions of the selected values. *NOTE: These values are not* $d1, d2, d3, d4$*; rather, they represent selected values.*

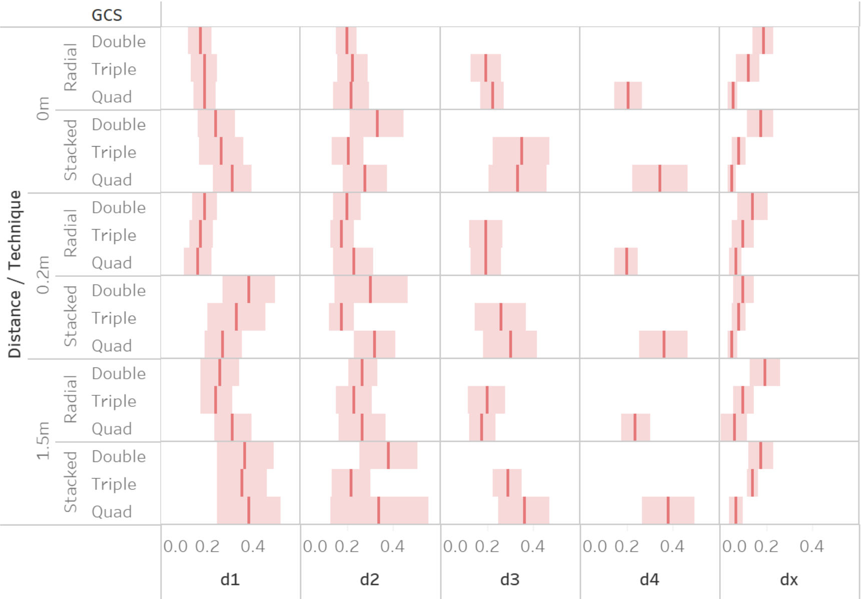


Figure 6.8: The medians of AbsDiff by distance, technique, and GCS with 95% confidence intervals generated using Tableau.

number of multiplications of unit interval numbers means the resulting value tends toward zero. Therefore, the participants simply chose zero for higher GCSs. Despite the overall inaccuracy, many of our tests still yield statistical significance.

Omnibus PERMANOVA Test

We performed an omnibus PERMANOVA test, and then posthoc PERMANOVA tests for statistically significant effects. Table 6.1 describes the results of all the tests.

	SS	**F**	**p**	R^2
Omnibus $(df = 1, df_{res} = 1719, \alpha = 0.05)$				
Technique	2.833	104.420	0.001*	0.046
Distance	0.280	10.320	0.001*	0.006
GCS	11.287	415.990	0.001*	0.183
MSD	0.159	5.860	0.002*	0.003
Tech. x Dist.	0.016	0.610	0.608	<0.001
Tech. x GCS	0.423	15.610	0.001*	0.007
Dist. x GCS	0.015	0.560	0.656	<0.001
Tech. x Dist x GCS	0.013	0.480	0.677	<0.001
Post-Hoc: Technique x GCS $(df = 1, df_{res} = 574, \alpha = 0.017)$				
Radial, Double v Stacked Double	2.771	62.602	0.001*	0.098
Radial, Triple v Stacked, Triple	1.424	40.042	0.001*	0.065
Radial, Quad v Stacked, Quad	1.874	48.743	0.001*	0.078
Post-Hoc: Technique $(df = 1, df_{res} = 574, \alpha = 0.05)$				
Stacked v Radial	2.833	83.117	0.001*	0.046
Post-Hoc: Distance $(df = 1, df_{res} = 1150, \alpha = 0.017)$				
0m v 0.2m	0.040	1.155	0.234	0.001
0m v 1.5m	0.205	5.655	0.001*	0.005
0.2m v 1.5m	0.250	6.833	0.001*	0.006
Post-Hoc: GCS $(df = 1, df_{res} = 1150, \alpha = 0.017)$				
Double v Triple	0.994	24.344	0.001*	0.021
Double v Quad	0.656	15.313	0.001*	0.013
Triple v Quad	0.682	17.384	0.001*	0.015

Table 6.1: PERMANOVA tests with the Bonferroni-adjusted post-hoc tests. * denotes statistical significance ($p \leq 0.05$).

The interaction between Technique and GCS was statistically significant ($F_{1,1719} = 15.540, p = 0.001, R^2 = 0.007$); therefore, we performed pairwise post-hoc tests. However, individual analyses of the main effects (Technique in Sec. 6.4.1, and GCS in Sec. 6.4.1) turned out to be more useful.

Techniques: H1.1

The post-hoc test on Technique was statistically significant ($F_{1,1719} = 104.150, p = 0.001, R^2 = 0.046$). Participants tended to perform worse with Stacked–meaning the result does not support H1.1. We observed that Stacked induced more head movement than Radial, as seen in Fig. 6.9. As such, we tested the MSD of the HoloLens's position. MSD was statistically significant ($F_{1,1719} = 5.860, p = 0.002$); however, the effect size was very small ($R^2 = 0.003$). Therefore, other causes must have been responsible.

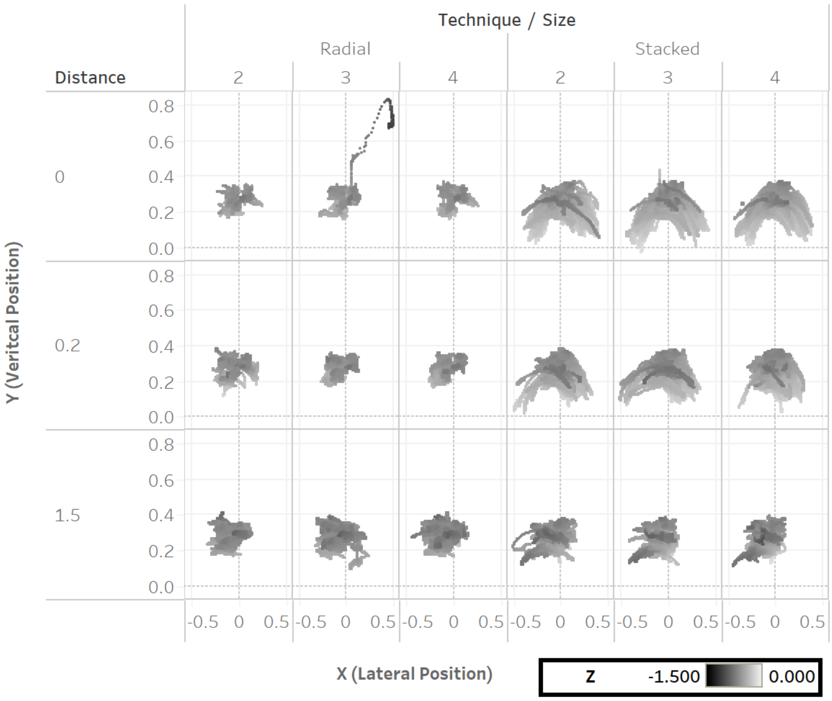


Figure 6.9: Distributions of the participants' HoloLens positions relative to the tablet. X, Y, Z represent lateral, vertical, and forward HoloLens positions respectively. The unit is metre.

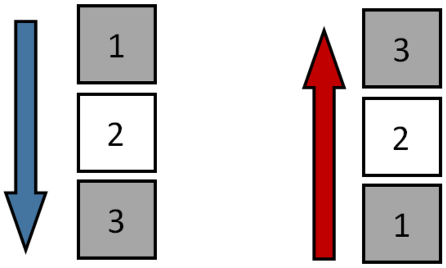


Figure 6.10: In a Triple Stacked glyph composite, the second glyph is always read the same way regardless of the reading order.

Glyph Composite Size: H1.2, H1.3

We found increasing GCS adversely affected both techniques ($F_{1,1719} = 415.990, p = 0.001$). This contradicted H1.2 which states that GCS does not affect accuracy for Stacked. This also means the ability to easily decompose glyphs did not improve the accuracy. Actually, the participants may have occasionally read Stacked glyphs in reverse order. Upon examining Fig. 6.8, we noticed that, on average, $d2$ for Triple Stacked glyphs was lower than the other Stacked AbsDiff. We realized: while all

Stacked glyphs would be wrongly assigned when read in reverse, the second Triple Stacked glyph would not. Due to it being in the sole middle glyph, it would be read as a second one regardless of the reading order. Fig. 6.10 illustrates this. The participants' pre-existing bias, possibly Spatial Numerical Association of Response Codes (SNARC) (see Shaki and Fischer [2018]), may be the root cause of this error. The training and Fig. 6.3 failed to totally eliminate this bias. Since Radial was also adversely affected by GCS, the test supports H1.3, which states that the colour of each constituent Radial glyph affected the perception of the other constituent glyphs.

Distance: H1.4

Distance was statistically significant, but its effect size was extremely small ($F_{1,1719} = 10.300, p = 0.001, R^2 = 0.006$) on $d1, d2, d3, d4$, and dx. The following pairwise tests were statistically significant: 0m v 1.5m ($F_{1,1719} = 5.655, p = 0.001$), and 0.2m v 1.5m ($F_{1,1719} = 6.833, p = 0.001$). The pairwise test for 0m and 0.2m was not. Although H1.4 was confirmed, other factors (e.g. GCS) had a much larger impact on AbsDiffs.

6.4.2 Duration: H2.1-H2.3

Fig. 6.11 shows the distributions of the duration of the trials. We performed an ART-ANOVA test (Table 6.2), then posthoc tests on significant effects (Table. 6.3). At 0m, both techniques were not statistically different in terms of duration, thus contradicting H2.1. Meanwhile, H2.2 was supported; Stacked was faster than Radial at 0.2m and 1.5m. Therefore, if the composites were further away, the ability to decompose glyph composites reduced trial durations. Interestingly, the pairwise tests on Distance did not support H2.3 since they showed that increasing Distance reduced trial durations instead of increasing them. It turned out, the participants spent less time on glyphs that they could not examine closely.

6.4.3 Usability and Body Movements: H3.1-H3.2

Fig. 6.12 shows the distributions of the UMUX scores. The median UMUX score for Radial was 66.667 out of 100, and the score for Stacked was 72.917 out of 100. Despite having different medians, the Wilcoxon signed-rank test comparing the UMUX distributions was not statistically significant ($W = 31.5, p = 0.929$). Therefore, we

	F	df	p	$\eta^2_{partial}$
Technique	0.005	1	0.941	≤ 0.001
Distance	9.701	2	$\leq 0.001*$	0.01
GCS	14.318	2	$\leq 0.001*$	0.14
Technique x Distance	2.257	2	$\leq 0.001*$	0.03
Technique x GCS	1.258	2	0.284	0.001
Distance x GCS	2.179	4	0.069	0.005
Technique x Distance x GCS	2.444	4	0.045*	0.006

Table 6.2: The results of the omnibus mixed effect ART-ANOVA tests on duration with the participant as the random effect. All residual degrees of freedom are: 1695. * denotes $p \leq 0.05$.

	Estimate	SE	df	t	p	d
Technique x Distance x GCS						
Stacked-Radial & 0m-0.2m & Double-Triple	-111.1	114	1695	-0.971	0.332	-0.024
Stacked-Radial & 0m-1.5m & Double-Triple	-157.4	114	1695	-1.375	0.169	-0.034
Stacked-Radial & 0.2m-1.5m & Double-Triple	-46.3	114	1695	-0.405	0.686	-0.010
Stacked-Radial & 0m-0.2m & Double-Quad	15.2	114	1695	0.133	0.894	0.003
Stacked-Radial & 0m-1.5m & Double-Quad	-274.8	114	1695	-2.4	0.017*	-0.059
Stacked-Radial & 0.2m-1.5m & Double-Quad	-290	114	1695	-2.533	0.011*	-0.062
Stacked-Radial & 0m-0.2m & Triple-Quad	126.3	114	1695	1.104	0.270	0.027
Stacked-Radial & 0m-1.5m & Triple-Quad	-117.3	114	1695	-1.025	0.306	-0.025
Stacked-Radial & 0.2m-1.5m & Triple-Quad	-243.6	114	1695	-2.129	0.033*	-0.052
Technique x Distance						
Stacked-Radial & 0m-0.2m	7.81	46.6	1695	0.168	0.867	0.004
Stacked-Radial & 0m-1.5m	274.92	46.6	1695	5.901	<0.001*	0.143
Stacked-Radial & 0.2m-1.5m	267.11	46.6	1695	5.733	<0.001*	0.139
Distance						
0m-0.2m	-47.5	23.4	1695	-2.033	0.105	-0.049
0m-1.5m	-102.8	23.4	1695	-4.401	<0.001*	-0.107
0.2m-1.5m	-55.3	23.4	1695	-2.367	0.047*	-0.057
GCS						
Double-Triple	-222	22.3	1695	-9.956	<0.001*	-0.242
Double-Quad	-375	22.3	1695	-16.828	<0.001*	-0.408
Triple-Quad	-153	22.3	1695	-6.872	<0.001*	-0.167

Table 6.3: The post-hoc tests (Tukey-adjusted) for the duration with Cohen's d. The unit is in milliseconds. * denotes $p \leq 0.05$.

were unable to prove H3.1. which states that Stacked is deemed as more usable.

We analyzed the answers to Q5 to Q8 which pertain to participants' movements. Fig. 6.12 shows the distributions of the participants' answers. The Wilcoxon signed-rank tests for Q5 ($W = 44.5, p = 0.692$), Q6 ($W = 23, p = 0.120$), and Q8 ($W = 31.5, p = 0.929$) were not statistically significant. However, the Wilcoxon signed-rank test for Q7 was ($W = 1.5, p = 0.005$)–albeit with a small Kerby's r Kerby [2014] of 0.011. This means that the participants believed they moved slightly more with Stacked which confirmed H3.2.

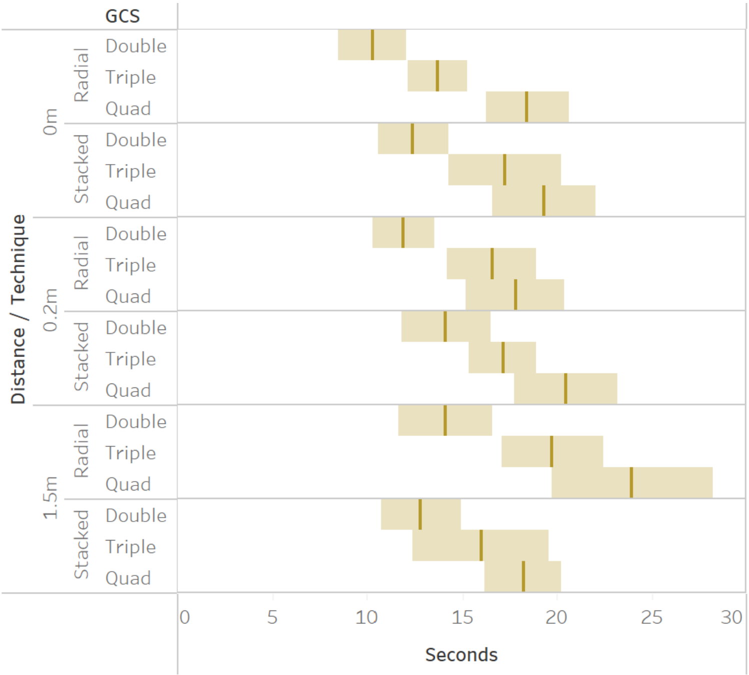


Figure 6.11: The medians of the trial durations by distance, technique, and GCS with 95% confidence intervals generated using Tableau.

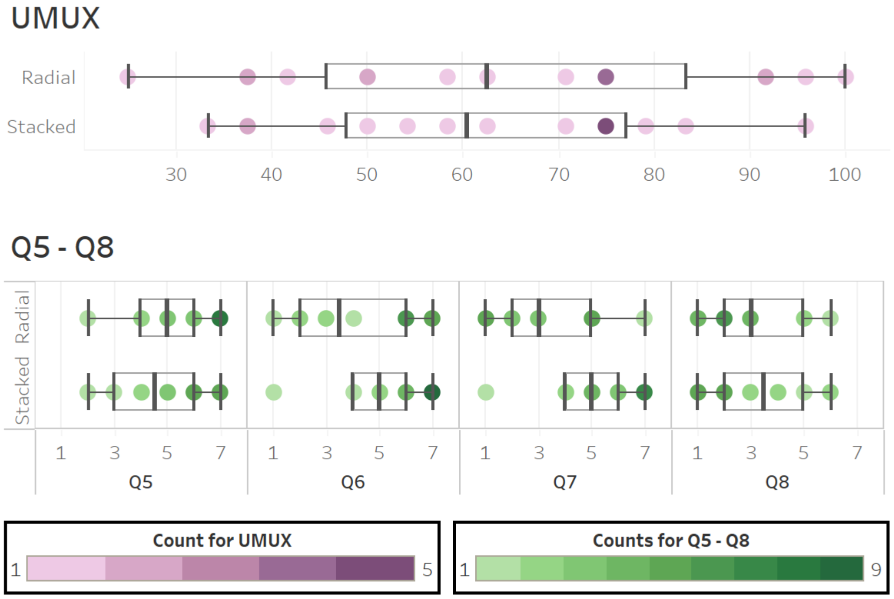


Figure 6.12: UMUX and questionnaire Scores. Each purple/green dot represents the frequency of a specific score.

6.5 Discussion

In this section, we disseminate and discuss the results of the study. Additionally, we identify possible future research directions to address the limitations of our study.

6.5.1 The Parallax Effect

The parallax effect did not have much impact on accuracy. Instead, GCS had more impact. Increasing the number of constituent glyphs in a composite makes it more difficult to interpret for both Radial and Stacked. However, more errors were found with Stacked. Furthermore, the visual aid (e.g., Fig. 6.3) and the participants' prior bias play an important role. We found that the participants may have read Stacked glyphs in reversed order, possibly due to their prior Spatial-numerical Association of Response Codes (SNARC) bias. SNARC explains how people link the orders of arranged objects with number [Shaki and Fischer, 2018]. There are many factors that influence SNARC, such as linguistic, psychological, and cultural [Göbel, 2015, Shaki and Fischer, 2018, Aulet et al., 2021]. Apparently, the participants' prior SNARC bias was so strong that neither the training nor the diagrams could overcome it. To verify if SNARC did indeed play a role, more research is needed.

On the other hand, the parallax effect has a clear positive effect on speed. It is the most beneficial when the glyphs are further away. In this case, the effect automatically decomposes a Stacked composite–making it easier and faster to examine constituent glyphs. Despite Stacked inducing more head movement, the tests on the trial durations did not show any difference between Stacked and Radial when the glyphs were close. This means the ability to easily decompose glyph composites may have compensated for the extra time spent on head movements. Additionally, while the participants felt that they moved more with Stacked, the effect seems minimal. Therefore, the extra movement may not have felt cumbersome.

6.5.2 Colourmap

Prior work [Crameri et al., 2020, Moreland, 2009, Stoelzle and Stein, 2021] recommends against using a rainbow colourmap; therefore, this study used a two-hue divergent colourmap. However, the participants ended up providing rather binary answers.

Counterintuitively then, a rainbow colourmap may have been more appropriate, because its multiple hues allow for dissecting the colourmap into smaller ones. This phenomenon was observed by Quinan et al. [2017], and Reda and Szafir [2021].

Additionally, we must reconsider the presentation of blended values. We are aware that as the GCS increases, the overlapping value tends toward zero due to having more multiplications of unit interval numbers. However, we thought the participants could still avoid providing a non-zero value. It turned out to still be too difficult for the participants. We propose two methods that may alleviate this issue. The first is using a different colourmap to represent the blended value. The colourmap's scale should allow lower values to be more easily distinguishable from each other. Another is taking the n-th root of value with n being the number of overlapping values. For example, let v_1, v_2, v_3, v_4 be the values, and v_m be the blended value, we visualize $v^*_m = \sqrt[4]{v_m}$. We note that since v_m is a multiplication of v_1, v_2, v_3, v_4, it is a hypervolume of a hyperratangle with the lengths of the four values. Computing v^*_m turns v_m into a length similar to the four values. We caution that these two methods may not be implementable as shader programs—unlike the method used in the study.

6.5.3 Limitations and Future Work

Since our study has an elementary design, it allows us to discover behaviours that synoptic studies cannot detect (e.g. the potential existence of the SNARC). While the obvious next step is to conduct studies with synoptic tasks (e.g. detecting correlation, identifying clusters Andrienko and Andrienko [2005]), our results also suggest other types of studies.

Colourmap Studies

Our study demonstrates that a divergent colourmap may not be appropriate. However, since our study did not have other types of colourmaps, we cannot conclude if the other colouramps could lead to higher accuracy. Additional studies are necessary. Furthermore, future work should explore novel ways to represent blended colour values as the current method of representation may not be effective. Lastly, our study excluded people with CVD. A future study should also include them, and identify the type of colourmap that is the most CVD-friendly.

Glyph Ordering Studies

We have potentially detected the SNARC effect for Stacked, and another effect for Radial. However, our study design does not allow us to disseminate these effects. We require at least two types of studies to fully understand these effects. First, there must be a SNARC study in 3D. As it turns out, all SNARC studies were conducted on flat surfaces. We propose a study that arranges objects in 3D using either a mobile sculpture or mixed reality. Secondly, we must explore the role of visual aid in understanding a glyph composite's inner structure; our study suggests that our own visual aid (Fig. 6.3) can be improved. Alternatively, a future study can sidestep SNARC by simply asking the participants to interact with a virtual replica of the glyphs and directly inputting answers through the AR interface itself.

Other Composition Techniques

In our study, we only explored using the parallax effect for composing and decomposing glyphs. However, other composition/decomposition methods are also possible. For instance, the user could use hand-based gestures instead. Future studies should compare such methods against Stacked.

Large Area Display and Navigation Studies

Since AR allows us to have a very large display area, we varied the distances to better understand how the user deals with faraway objects in such a large area display. Having only three levels of distances is helpful for keeping the number of trials manageable. However, we can envision a future study where levels are continuous, and not fixed like in this study. The result of such a study will allow us to better understand how glyph decomposition can speed up the user at a continuous scale.

We can also explore how search directions affect effectiveness in the future. Instead of only asking the participants to search on the left side like in this study, a future study should ask the participants to search in all directions. The study will require us to rethink the visual aid as well. In our study, we only used a left-pointing arrow which is insufficient for future study. This necessitates additional research in cueing techniques for out-of-view glyphs (e.g. Burigat et al. [2006]).

We restricted our participants' movements which eliminated navigation as a confounding variable. This allowed us to understand their perception of far-away composites in a large area display. However, for the user, having glyphs so far away may not feel comfortable to them. Therefore, some users might use navigation (e.g. walking up to the glyphs) to deal with the issue. A future study can relax the movement restriction to observe how participants can use navigation to deal with distance.

Additionally, instead of dealing with a single composite, the participants could perform tasks with a glyph field like in Chapter 5. Such a study will further demonstrate how the results of this study can scale up. More sensitive questionnaires will be necessary, however, considering the questionnaire in Chapter 5 is not effective.

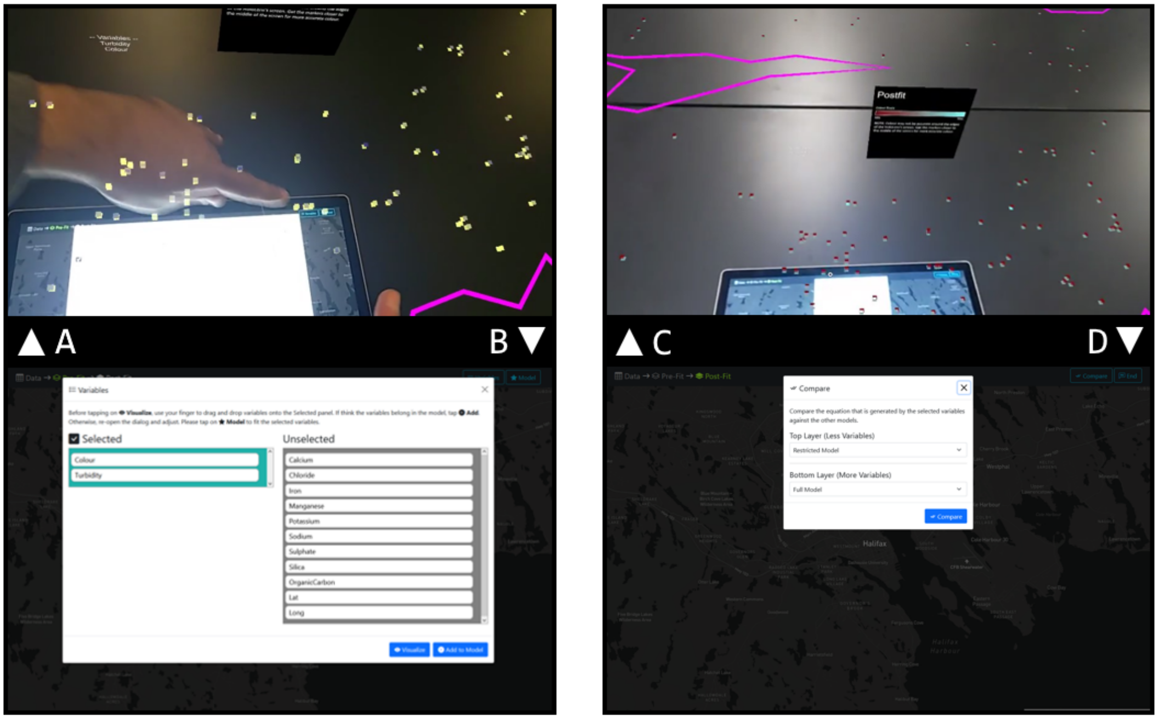
6.6 Conclusion

The synoptic study reveals that the dimensionality of the composite does not have much impact on comprehension. Rather, other factors such as the inner structures of the glyph composites play a much larger role. We also found the divergent colourmaps may be ineffective. However, more work is necessary. While the study allows us to better understand various elements of glyph comprehension, it also shows that there are many other challenges. Addressing these challenges should yield a more effective effect size questionnaire (ESQ). However, trying to address them requires working with the field of psychometrics, a topic beyond the scope of this thesis.

The walkthrough demonstration study, described in the next chapter, does not aim to address the new challenges identified in this chapter. Rather, it focuses on obtaining expert feedback. The synoptic study, as we have previously stated in Ch. 5, aims to fulfill all the research goals (Obj1, Obj2, Obj3). However, while the study yields some insights on how on a user may interact with Gander during the pre-fit and post-fit stages, the study design is too controlled. This means the study lacks useful qualitative feedback for improving the prototype. Addressing this gap is paramount to our work.

Chapter 7

The Walkthrough Demonstration Study: An Expert Evaluation of Gander



Before tapping on Visualize, use your finger to drag and drop variables onto the Selected panel. If think the variables belong in the model, tap Add. Otherwise, re-open the dialog and adjust. Please tap on Model to fit the selected variables.

Compare the equation that is generated by the selected variables against the other models.

Top Layer (Less Variables): Restricted Model

Bottom Layer (More Variables): Full Model

Figure 7.1: A: An investigator guiding A2 through Gander in the pre-fit stage looking at the relationship between two variables. The purple line indicates the map boundary. The colourmap is Ukraine. **B:** The reconstruction of the tablet interface shown in A. **C:** A2 using Gander in the post-fit stage to understand the distributions of the likelihood. The purple line indicates is the map boundary. The colourmap is Chicago. **D:** The reconstruction of the tablet interface shown in C.

The walkthrough demonstration study is for fulfilling Obj1. The focus is to understand how an expert might approach Gander, and to use their feedback to improve Gander. Although the synoptic study also evaluated Gander, the experimental control makes generalizing the results into tangible design improvement difficult. Unlike the elementary study which focuses on glyph comprehension (Obj2), this study does not analyze the effectiveness of the visualization technique. Rather, the study targets the whole system.

We evaluated the prototype described in Ch. 4 using interviews and walkthrough demonstrations. A walkthrough demonstration involves the investigator guiding the user through the interface [Ledo et al., 2018]. An example of this is the work by Evangelista Belo et al. [2021]. They developed software that allows VR designers to create ergonomic VR software. They invited experts and guided them through a tool instead of allowing them to explore the tool themselves since the tool has some learning curves. A walkthrough demonstration is appropriate here because Gander has a linear workflow with a specific end goal. Furthermore, our pilot studies indicated that the interface can be difficult to use at first. Therefore, some guidance may be necessary. This approach is appropriate in this case because Gander is a vertical slice prototype. As a vertical slice, Gander allows the user to complete the fundamental IML tasks (i.e., data selection, data exploration, model fitting, and model assessment). However, its functionalities are limited–e.g. the user can only perform MLR. Due to the lack of options, the participants were effectively "railroaded." Had the participants been allowed to use Gander without any sort of guidance, they would quickly encounter limitations and become frustrated.

Six participants from various backgrounds (i.e. maritime, usability analysis ecology, and agriculture) participated in the study. They practice various types of spatial analysis at the professional level. The recruitment process was extremely challenging due to the potential participants not being available due to their jobs. Many potential participants also worked remotely and did not have a consistent Internet connection. As such, we must slightly modify the procedure to suit their availability and circumstances. In general, however, we first interviewed the participants for their background knowledge in spatial analysis. Then, we asked them to interact with Gander. Finally, we interviewed the participants for the second time on their experience with Gander.

7.1 Participants

We recruited the following experts: (1) two geospatial analysts working with the government of Canada, (2) two usability experts, and (3) two lecturers working at rural universities. The first group of participants worked with maritime geospatial data and their work had influenced on Canadian maritime traffic. We assigned the participants in this group the following ID: A1, A2. The second group was not geospatial analysts; however, they evaluated geospatial analytics software used by the first group. They were assigned: B1, B2. Their evaluation also involved mixed reality prototypes. Therefore, we deemed their insights as valuable. The third group of the participants possessed the most technical backgrounds. Not only they advised people on conducting research with geospatial data, they also taught advanced classes. We gave them the following IDs: C1 and C2. Please refer to Section 7.4 for the full descriptions of the participants.

7.2 Procedure

In the first interview, we performed a semi-structured interview where we asked the participants about their backgrounds. We asked the participants slightly different questions based on probing and how the participant responded. The original questions were:

- Can you describe your background and your work with geospatial analysis?
- What are the main processes in your work and can you describe the datasets that you use?
- Can you walk me through a typical process for geospatial analysis?
- What are your primary spatial analysis techniques?
- What details will you be looking for during the analysis?
- What are the challenges with your analysis?
- How do you communicate your analysis to other people?
- Are you familiar with augmented reality or virtual reality? What is your impression of the technology?
- Is there anything else you would like to say?

During the walkthrough demonstration, we displayed a short video explaining the inner workings of Gander. The data used in the walkthrough demonstrations were modified from Nova Scotia Lake Chemistry data The Government of Nova Scotia [2021]. Furthermore, the particular data have minimal overdrawing. In the exit interview, we asked the participants for their feedback. The original questions were:

- What do you think about the demonstration?
- How well can you understand the interface?
- How can this work be improved?
- How can Gander be incorporated in your line of work?
- Do you think AR can enhance spatial analysis?
- Do you have any additional comments?

Given each participant had different availability, tailoring the procedure to suit their schedule was necessary. A1, A2, B1, and B2 each attended a 30-minute session that included a background interview plus the walkthrough demonstration and an exit feedback interview. The 30-minute was hybrid—i.e. the participant interacted with the prototype in person; however, they were interviewed by a remote interviewer through a computer. A1 agreed to an additional remote interview. In this extra session, we asked the A1 questions on the communication of his results, the stakeholders involved in the process, and the post-fit stage.

Before the interview, C1 requested for the training video. Afterwards, C1 attended a 45-minute session of a background interview plus the walkthrough demonstration with an exit interview. Because C1 was unable to trial Gander in-person, the walkthrough demonstration involved C1 remotely reviewing the video and and critiquing the interface presetned in the video. The lack of in-person interaction could be seen as beneficial, because C1 would be less affected by the novelty effect. Although C1 had more time than A1, A2, B1, and B2, the questions did not change based on the extra time. Instead, C1 received more time to ponder and answer the questions.

We performed a 60-minute with C2 in-person at their office. Their procedure was the same with A1, A2, B1, and B2. Like C1, the extra did not result in a significant change of the procedure. He simply had more time to respond and to interact with Gander.

7.2.1 A Note about the Overlapping Areas between Glyphs

The participants were not made aware that the overlapping areas between the glyphs are multiplied in the pre-fit stage, and then subtracted during the post-fit stage. We made this decision in light of the elementary study (Ch. 6) which has demonstrated that an individual tends to have issues gleaning information from overlapping areas.

7.3 Analysis

7.3.1 Thick Description

To analyze the interview and the demonstration walkthrough, we rely on thick descriptions. To explain what constitutes a thick description, we must first turn toward how human-computer interaction (HCI) treats qualitative analysis. Typically, HCI relies on thematic analysis. With this approach, common themes found in individual interview data are extracted, reduced, and organized [Braun and Clarke, 2006]. The goal is to find commonality.

A thick description, first introduced by Clifford [1973], also seeks to understand the common themes. However, unlike thematic analysis, it provides vivid details of the participants. For instance, instead of simply stating that the participants stated one similar thought in the interviews plus an overview explanation, thick descriptions provide one or more stories of how the participants arrived at the similar thought. For instance, in this walkthrough demonstration study, we state each participant's position and the tasks that they perform in their professional capacity, so everyone can understand how the participants thought about Gander. Based on the work of Kharel [2015], we think a thick description should: (1) provide the context of the study data, (2) outline the meaning behind the data, (3) tell the circumstances that give rise to the data, and (4) be interpretable as text. A thick description is vivid, highly detailed, and very descriptive. Porter [2012] states that a thick description does not simply mean a highly detailed description that includes every single minute detail. Rather, the focus is to extract meaningful richness from the data.

The main benefits of using thick descriptions are: gaining a deeper understanding of individual users, and understanding why individuals may think in certain ways. An opposite of a thick description is a thin description [Clifford, 1973]. A thin description provides a surface-level description, and lacks motivation. Porter [2012] argues that some thin descriptions are thin not by choice, but by necessity. For instance, modern computer scientists may no longer identify with early programmers who worked with vacuum tubes and punch cards, because the hardware and the experts are difficult to find. Therefore, research involving early computers may be replete with thin descriptions. Other examples of thin descriptions include statistical analyses which mainly focus on establishing statistical validity over providing individual explanation [Porter, 2012]. An example of this is a Fitts's law model which predicts the speed of target acquisition from an origin. The model does not explain why and how the speed would be achieved. A thin description is not necessarily a useless one. In fact, both types of descriptions can work together. As Brekhus et al. [2005] argues, thin descriptions can be used to provide impersonal perspectives at the beginning, and to provide general overview information. Then, the thin descriptions can be used to motivate further scrutiny of the data which yields thick descriptions. We also think forcing a thin description to become a thick one is against the core tenet of science as it involves embellishment or imagination that is ungrounded and unsupported by science. We are able to generate thick descriptions in this study, because our interviews were primarily about gathering the participants' background information. This is different from the synoptic study where the participants were merely comparing two visualization techniques.

While thick descriptions are not standard in HCI, there are several examples. In the work of Nas et al. [2023], the participants answered an online questionnaire. In each question, each participant saw an image, representing how people in the past saw future technology. They can then submit their own image in response. In addition to the image, they submitted a thick description of themselves. Another work by Cheung et al. [2014] contains a story of how a new video player can become engaged and disengaged during the first hour of gameplay. The formats of thick descriptions can vary. For instance, the original example of thick description in Clifford [1973] reads like a fiction, the ones in Nas et al. [2023] read more like personal statements, and Cheung et al. [2014] created a single description from the perspective of a single generic user. Our thick descriptions somewhat resemble the ones shown in Nas et al. [2023]. Each description is from a single participant, and it outlines the participant's background information and thoughts. However, unlike Nas et al. [2023]'s description, our descriptions do not contain any element of positionality statements–i.e. we did not discuss the participant's lifelong experience and its impact on their thinking. Due to the diverse backgrounds possessed by the participants and the small sample size, we are unable to create a singular thick description representing a generic experience like in Cheung et al. [2014].

For each participant, we first describe how the procedure was modified to suit their availability. Then, we summarized their background. Afterwards, we discussed their individual feedback on Gander. Finally, we provide a summary of all participants' thick descriptions.

7.3.2 Other Methods

In addition to thick descriptions, we use the qualitative analysis method outlined in Reilly and MacKay [2013]. In their work, they interviewed biologists to understand how they annotated ecological data collected from fieldwork. Before interviewing the participants, they performed a first principle analysis–that is, trying to understand what constitutes the "standard practice" in the existing literature. Then, through the interviews, they identified the differences between the first principle analysis and the actual practices. Gander was designed from the perspective of the first principle analysis. While we evaluated certain aspects of Gander through the elementary, and the synoptic studies, the studies are highly quantitative in nature. As such, prior to this study, the system still lacked qualitative feedback from actual experts. Therefore, analyzing the interview data from the walkthrough demonstration allows us to understand how to improve the prototype. We also used the "bottom-up" thematic analysis as proposed by Braun and Clarke [2006], Bruan and Clarke [2012] to summarize the thick descriptions.

7.4 Results

We separate our results into three parts: thick descriptions, the common Gander Workflow, and the Walkthrough Demonstration. The first one provides the thick descriptions that we have created based on the interviews. They have been formatted similar to the ones found in Nas et al. [2023]. Additionally, they outline how we adapted the procedure to suit the participants' availabilities. The second part describes the steps and procedures deployed by the participants when working in their capacity. The last part is the summary of the feedback of Gander provided by the participants.

7.4.1 Thick Descriptions

In addition to the thick descriptions provided below, we provide a summary table (Table 7.1) which contains the following information: position, type of data, technique used, software used, experience with MR, and sessions attended.

	A1	**A2**	**B1**	**B2**	**C1**	**C2**
Position	Spatial Analyst	Spatial Analyst	Usability Evaluator	Usability Evaluator	Researcher, Lecturer	Researcher, Lecturer
Spatial Data	Bathymetry acoustic data, shape data	Ship trajectory data	None	Gaze heatmap data	Remote sensing data, rural data, ecological data	Remote sensing data, land use data
Technique Used	Maritime acoustic model, probabilistic risk model for maritime life	Descriptive maritime trajectory analysis, Pattern-of-Life analysis, temporal analysis	Statistical inference	Descriptive trajectory analysis	Kriging, interpolation, GWR, spatial regression, covariance modeling	Spatial regression, Classification
Software	Python (Shapely, GeoPanda), QGIS	ERSI, QGIS	None	None	Web-based apps provided by the data sources, ArcGIS, Google Earth	GIS, ENVI, PCI Geomantica, R
Prior MR Experience	Hackathon	VR digital twin mockup of marine vehicles	Conducting AR studies	Conducting AR studies	None	None
Session Time	30 min + 20 min follow-up	30 min	30 min	30 min	45 min	60 min

Table 7.1: Summaries of the participants, their background, and session information.

Thick Description 1: A1

The study was conducted in a hybrid format. The principal investigator interviewed the participant remotely through a computer while the other investigator conducted an in-person walkthrough demonstration. Several weeks later, the principal investigator interviewed the participant remotely.

Based on the interview, we found A1 to be enthusiastic about his work. We learned that A1 did not have any formal education in geospatial analysis. Instead, he was trained as a mathematician with an additional background in finance. His expertise was in generating mathematical models. He had knowledge of statistical modelling and time series analysis. A1 was only exposed to MR through a Hackathon. As a part of his work, he was creating a probabilistic risk model for marine life. He relied on bathysphere data, and acoustic data to generate his model.

For the software, he stated that he exclusively used GeoPandas, a Python library. He primarily worked with shape files. In his current project, he was creating a risk model for marine life. Ultimately, the government could use the model to instruct ships and marine vehicles to avoid certain areas in the sea in order to protect marine lives in those areas. He had experience using Python and rarely relied on data visualization. Geospatial visual analytics tools (e.g. QGIS) were only used to gain some familiarity with the data set. However, once A1 became familiar with the data, no more visualization was used. He did not use any geographic information system (GIS) at all for the current project that he was working on.

After the walkthrough demonstration, A1 stated that Gander was: "pretty intuitive". However, he struggled somewhat with the dialog boxes, as he stated: "Only part that I would be a little confused by is having to like close out of menus (dialog boxes) to see things (glyphs)." When asked if Gander could help with his line of work, his replied that it could. He said "I think the advantage is you can easily visualize multiple layers at once over a pretty large area, and because I know if you were looking at geospatial data over some large region." Furthermore, the ability to navigate through the glyph field was a bonus: "You might want to zoom in on certain things and the augmented reality lets you zoom in and move around." We must note that *zoom* here did not refer to the unimplemented zooming function. Rather, it referred to the ability of the user to get spatially close to the glyphs. A1 also made a comparison between Gander, Python, and MATLAB: “Python it’s very difficult to do that (visualization). I know like in MATLAB it’s much easier to rotate and zoom around in your plots, but in Python not so much. So that’s definitely a nice feature that I like is being able to like, sit down and look at output from your models and be able to like figure out what it is that it’s showing you.” The comparison showed that A1 wishes that he could be able to check statistical information at a granular level.

During the walkthrough demonstration, we overheard A1 stating that contours could be used instead of glyphs. Therefore, after the demonstration, we asked A1 if he would want to replace the glyph-based system or not. He stated that the contour would be better for the post-fit stage: “I think [contours] might be helpful [for] relative likelihood. Then you could then you could look at the relative; if there were two models there, you could look at the relative likelihood between them over the whole map. So then you’d know the regions or the model is fitting better than other ones. It would be much more clear.” However, it would not be as useful during the pre-fit stage, because “if you have multiple variables or multiple models; you wouldn’t really be able to do that, so yeah...” These statements suggested that glyph-based visualization was less helpful for indicating spatial autocorrelation. However, it could be used when there were too many dimensions.

The purpose of the follow-up interview was to further understand A1’s work process. During the previous interview, we learned that he was creating a probabilistic risk model for marine life. However, there was no specification on what the model could be used for. He clarified that although he created a model, there were other decision-makers who: “would be using information on the number of marine mammals that are at risk of being harmed.” He stated that he visualized the output of the model using a raster over a map, and that the raster was for other stakeholders. He also explained that his model was created using existing statistical models: “... calculations I was doing are making use of models that other people have. I calibrate it to real data and I’m just using their model output in order to calculate something with it...” We also learned that A1 did not make decisions based on his model. Instead, his colleagues interpreted his model and made decisions: “Does that mean you stay away from that area or like what value for the risk tells you what you should do, and that’s someone else’s job to figure out how to do that. All I had to do was figure out how to assign risk to locations in space.” The model that he created was used to control maritime traffic in a way that ensured the risk to marine life was minimized. The main lesson that we learned was that statistical modelling always had multiple stakeholders, and sometimes, the non-users were actually the ones who were the most affected. If A1's model turned out to be wrong, the maritime operators would be paying for the cost.

Thick Description 2: A2

A2 was a colleague of A1. We learned that he had an extensive background in geospatial analysis. He stated that his work was related to the ocean. He used the following types of data: bathymetry, sea surface temperature, and shipping trajectories. Although he did not create any statistical model, he used a technique called Pattern-of-life analysis. Pattern-of-life analysis is a data mining technique that aims to extract the common, habitual, and repetitive trajectories of entities Grégoire [2013]. In A2's words, he described his own pattern-of-life analysis as: "the ship traffic, if I draw an analogy to between sort of vehicle traffic, vehicle traffic has certain patterns that route that they follow, but also temporal patterns or for example, rush hour has more cars than midnight traffic. So the patterns that the traffic follows in terms of space, but also in terms of time and that goes into. The overall statistics of the patterns of the vessels on the water. That's pattern of life." Although A2 was not using statistical modelling at this moment, he stated that he relied on statistical methods in the past.

The participant stated that he used GIS software. In general, A2 did not encounter any issue with his current GIS software. He used Esri ArcGIS in the past, but switched to QGIS. He stated that QGIS, as free software, was easier to procure.

To present the results of pattern-of-life analysis, A2 generated images that outline the patterns. Then he could present them using PowerPoint presentations or print them on paper.

A2 stated that he was always familiar with MR software, and had been using it for five to eight years at this point. The software was used in multiple ways. First, MR was used to create virtual immersive mock-ups that A1 could personally explore. Such mockups allowed for preliminary inspection without having to build a physical prototype. Secondly, MR was used to visualize traffic trajectory data. Thirdly, MR was used to visualize underwater space.

After the walkthrough demonstration, A2 expressed that the interface “was very good” because “It seemed clear and uncluttered.” When asked to define “clear”, A2 said that “clear” meant that “Clear means I could ascertain what I was going to do.” For “uncluttered”, he defined it as “uncluttered means it didn’t have extra options for my actions.” This meant, overall, according to A2, Gander had a streamlined interface that allowed him to accomplish his tasks without much distraction. However, he later stated that for Gander to be suitable for his line of work: “That would require a bit of thought.” He found the tablet-based menu needed improvements–particularly for the buttons which should have been more colourful.

Thick Description 3: B1

B1 was a cognitive psychologist. Although she was not a spatial analyst, she evaluated AR-based software used for geospatial analysis used by A1, and A2. As such, she still had valuable information for Gander. B1’s work included evaluating a geospatial analysis application which combined AR with a touch-enabled table-top. The user of the application performed touch gestures on the table-top display to query maritime data.

She did not perform spatial statistical analysis. Instead, she typically performed parametric and nonparametric tests with the data collected from the evaluation studies. Occasionally, she performed regression analyses. In addition to quantitative analyses, she also reported qualitative data. She was not involved in the design process of the AR and the table-top displays. Instead, designers used the quantitative and qualitative analyses performed by B1 to update the design.

The evaluation led to some changes in the design. An example used by B1 was how the user could measure distances between two vessels. In a previous design, the user must tap two visual representations of the vessels in order to obtain the measurement. However, in the updated version, the user could drag a vessel on top of another to get a pop-up that states the distance between them.

B1 had experience with MR through evaluations of MR systems. Therefore, her feedback from the walkthrough demonstration was valuable. She stated that Gander was: “pretty intuitive and straightforward.” However, she did not know if the AR interface could be beneficial. She stated: “I question whether the AR immersion is really adding much to it. I was just looking at the whole Nova Scotia map the whole time so that could have just as easily been shown on a tablet or a Laptop screen." She was also uncertain if the 3D layer approach was useful or not. However, she thought that the tablet interface was fine for the purpose.

Thick Description 4: B2

B2 was a usability evaluator. She indicated that she evaluated and developed an AR + tabletop system for geospatial analysis. She clarified that she did not use the tool herself. Instead, the agency did. Although she was not a geospatial analyst, she was aware of the requirements for a geospatial tool, as she put it: "I mean I'm experienced in understanding what geospatial data requirements are ... for [the agency] in order to make decisions. So I would say that [the agency] is the user of the geospatial data and I'm doing the user interface design for a tool that supplies that information to [the agency]." She further elaborated that the system was for oceanic data analysis. The data included marine acoustic data, temperature data, and acoustic data.

Although she was not a geospatial analyst, she analyzed spatial data in the form of eye-tracking data. She used eye-tracking data to identify where her participants were looking when performing tasks related to the evaluation of the AR + tabletop interface. Her goal was to understand where the user would be looking at when performing a task. She did not use any statistical analysis technique.

We learned that the agency had been evaluating MR interfaces for years. As such, B2 was quite well-versed with MR technologies and was familiar with it, as B2 stated: "Well, we've been working with augmented and virtual reality for ... I don't know ... a number of years. ... We got the HoloLens when it first came out, so we were doing concept development with it for a while. At this point, we're doing something that may be similar to what you're going to show me when using augmented reality to supplement a tabletop display."

After the walkthrough demonstration, B2 stated that Gander was intuitive: "I thought the interface was like the setting it up was very intuitive." However, she found the general glyphs' positions to be too close, as she stated: "I think that from like for the air stuff like. I mean, just generally like where it's positioned is awkward. And maybe if there was a little more difference between the heights of the two colours, it would be a little bit more obvious." She thought the user should have the ability to raise all glyphs from the tablet if the user feels the glyphs are too low. When the user is standing too tall, they can have difficulties understanding the glyphs–as B2 said: "I was standing up straight, I didn't really understand [the layers] I was trying to look at." She also wished that she could directly manipulate the glyphs, instead of "needing to go back [to the tablet] and reset [the visualization]." B2 added that if the user was not familiar with the data they were working it, they would need to constantly be relying on the tablet: "I guess if you don't really understand the domain, then it's gonna be a bit of trial and error."

Thick Description 5: C1

C1 was a researcher and a lecturer at a rural university. His expertise was in geoscience and hydrology. He had experience using GIS software, and remote sensing data. Examples of remote sensing data included Landsat data, field-collected data points, ground station-based metrological data, and more. He created statistical models for soil erosion, sedimentation, surface energy balance, and possibly more. His techniques included analysis of covariance, interpolation, and kriging. Speaking to him, it was clear that he had an expansive background. He stated that he chose his technique based on the target audience, and the data. For the target audience, he said that he chose simpler techniques if he was teaching beginners or performing casual research. He stated: "If I'm teaching the lower-level GIS courses or doing some casual research without any publication goal or anything, I would [perform] simple ... statistics." C1 would use more advanced techniques for publication, and teaching advanced students.

C1 indicated that web-based data portals had substantial capabilities for data visualization despite not being as advanced as a GIS like Google Earth, ArcGIS Online, and ArcGIS Pro. He stated that: "There are so many web applications developed at this time. For example, if you wanted to do some of the satellite analysis, The Landsat has its own [interface where] you can ... visualize and do more analyses. Sentinel, the European satellite, has its own web application. So do the local governments and state governments Local governments and agencies have started developing their own mapping platform." Ultimately, he stated that various tools were for data visualization, and data exploration.

C1 did not have much experience with MR. He explained it was why he asked for a video demonstration of Gander before the interview. He did not understand the video demonstration, so we replayed the video and slowly explained the workings of the prototype. After he fully understood the prototype, we proceeded with the post-demonstration interview.

During the interview, he stated that he found the transparent background was confusing. He stated that: "[transparent background] to visualize the data point when you project that, on the floor or any object in the room and that–I did not find that very helpful." He added: "I tried to ignore what is in the background and just focus on the data." He preferred that the AR content be projected onto a wall on a flat floor, as he stated: "Projecting and displaying it on an even surface would make it more consistent."

Thick Description 6: C2

C2 was a lecturer at a rural university (not the same one as C1). He stated that he taught remote sensing, GIS, and spatial statistics. These techniques were also used in his research. He provided advice to many entities–from students at the university, to the government. He clarified that he did not have much industrial experience. He usually worked with rural ecological, and agricultural data.

When dealing with raw geospatial data, he stated that they must first be pre-processed. He stated: "What mentioned in the project design ... Basically, so of course, [data] made from different sources, [we must] trim the data. [For each project, things] are different, ... because there will be scale issue, [data] may be very large, and you want to trim or clip to study [for specific] area, and the data may [need to] be normalized or standardized." Data must be pre-processed before they become usable in an analysis. C2 also elaborated that often, we must combine data from multiple sources.

Sometimes, surrogate or artificially produced data must be used because the real data were missing or severely affected by climate change. Occasionally, data for the whole region of the study were unavailable. Therefore, analysis could only be performed at the local level. C1 cautioned that the analysis performed at the local might not be applicable to the whole region: "The data from one region might not represent the pattern in the whole area, but it's still giving us some clue about what [is] going on or what happened in the past."

After the data were processed, he would perform classification, or hotspot analysis. Although not explicitly stated, it was implied that exploratory data analysis was used. For spatial analysis software, C2 stated that one could use ENVI, PCI Geomantics, or R.

To communicate the results, C2 created multiple types of communication: "tables, pictures maps, the charts, chart animations." He clarified that a chart animation is an animation image file that shows changes from one period to another.

When asked about his experience with AR, he stated that he was aware that it was about combining virtual content with real-world content.

We asked C2 about the future developments of GIS. C2 replied that he was not able to provide the full answer, he said that cloud computing and AI are changing the industries. As an example of how cloud computing and AI are changing GIS practices, he explained while the current online GIS are limited in terms of functionalities, the systems could actually deduce the appropriate analysis techniques to the user–as he put it: "From my perspective, ... it has very limited functionalities, but in some ways, [online GIS are] very smart; you just fill it with data and the app starts think about the type of data you have and then provide you with some options or solutions. And in contrast, [to use] the desktop version, you still need to know more about the different types of maps, and more about how to create [them]. So in the future, many of the GIS tasks will be highly streamlined. They will be much easier to handle and in many cases especially those tools to create the GPS apps for people to use. Many GIS analyses [will no longer need] GIS professionals."

After the walkthrough demonstration, C2 implied the use of 3D visualization in Gander was quite limited. However, it could become a more affordable alternative to data physicalization: "It should be used as like a cheaper version of 3D printing, like if you want to print something out. If you don't want to put in something else, but you still want something in 3D, then augmented reality can help with that." He stated that Gander could be used for teaching statistics. He asked: "Is it possible that a prof. who teaches the statistics course for the student? – Like, introducing your app or your tool or incorporate your tool in teaching ... either remotely or in class."

7.4.2 The Common Gander Workflow

Design as research was used to identify the workflow that geospatial analysts may use; each step in Gander was supposed to represent how a user operates. However, the interviews reveal that Gander may be missing certain steps. First, some experts identify the end goal before beginning any work. C1 indicated that geospatial analysts may modify their procedures based on the end goal. He stated: “If I’m teaching the lower-level GIS courses or doing some casual research without any publication goal or anything, I would [perform] simple statistics.” Secondly, the user may involve other stakeholders in the decision-making process. For instance, A1 stated that he deferred to his colleagues for interpreting his model’s outputs. He stated: “[It is] someone else’s job to figure out how to [interpret the risk]. All I had to do was figure out how to assign risk to locations in space.” This highlights the need for collaboration between the user and non-users.

Geospatial software packages can vary in terms of functionalities. Some software, while more limited in terms of features, is more automated. This can affect the workflow, as C2 stated: “From my perspective, [online GIS] has very limited functionalities, but in some ways, they are very smart; you just fill it with data and the app starts to think about the type of data you have and then provide you with some options or solutions. In contrast, [to use] the desktop version, you still need to know more about the different types of maps, and more about how to create [them].” C2 further argued that online GIS software could democratize geospatial analysis to those without prior technical experience. These users need a workflow different from the one we identified in our design process.

We did not hear much from the experts about the post-fit stage. From listening to the participants, we found the post-fit tasks are usually performed only with aggregated overview information (i.e. test statistics). Furthermore, some experts would not benefit from the post-fit stage at all; for instance, A2 and B2 only worked with descriptive statistics, and would never fit a mathematical model. A1 provided additional thoughts during the follow-up interview. He believed that visualizing the Akaike Information Criterion (AIC) would be better than visualizing Johnston et al. [2006]’s effect size, because more than two models could be compared at once.

7.4.3 Walkthrough Demonstration

Applying the "bottom-to-top" approach of thematic analysis (see Braun and Clarke [2006], Bruan and Clarke [2012]) to the thick descriptions, we categorized the post-walkthrough feedback into the following themes.

Overall Feedback. In general, the participants deemed the interface as good due to its general ease of use, and streamlined appearance–as A2 stated: "[Gander] was clear, and uncluttered." B1 said commented that Gander was "pretty intuitive and straightforward." However, both participants also had some criticisms: A2 thought Gander was too general for his tasks, and B1 commented that the AR interface could be replaced with a desktop-, or a laptop-based one. Furthermore, we must specify that although Gander was deemed easy to use in general, the participants still experienced difficulties.

Improvement for Tablet. We noted that several experts (A1, A2, B1) wanted more improvement from the tablet interface. They still found the tablet interface too cumbersome to use. An example of this is when the participants tried to change to the pre-fit visualizations. To do so, they must relaunch the Variable Picker, and re-select the variables–as B1 summarized it as: "needing to go back [to the tablet] and reset." Therefore, a future version of Gander should increase its support for direction manipulation, which has three elements: (1) continuous graphical representation of objects, (2) physical actions instead of complex syntax, and (3) the results of actions (including undoing action) are instantly graphically updated [Hutchins et al., 1985]. While Gander supports aspects of direct manipulation, required actions to update the visualization must be simplified and more accessible.

Improvement for AR. C1 thought the AR map's background should not have been transparent. He explained that he must "try to ignore what is in the background and just focus on the data." He added that AR content should be mounted on other surfaces rather than on the tablet. A1 and C2 indicated that the ability to zoom in and out is also important. C2 stated: "I did not get to see the whole map at one time." C2 pointed out that walking away from the map and glyphs could simulate zooming, because they would appear smaller–i.e. change in angular size. However, the ability to change the zoom level via the interface would be better. Lastly, A1 wished that there could be an alternative to the glyph-based visualization, as he was working with shape-based data instead of point-like ones.

Overview Information. C1 and C2 stated that aggregated information and statistics were vital for statistical inference. An example of aggregated statistics is a statistical table outlining a MLR model's test results. C2 stated: "[Gander should] just show some [overview] results. Or [it] may show some charts or figures." Glyph-based visualizations alone were insufficient.

Design Appropriation. C2 suggested that Gander could be used for pedagogical purposes. He said that we should be "introducing your app or your tool or incorporating your tool in teaching ... either remotely or in class." By proposing a novel and unintended use of Gander, C2 introduced the concept of design appropriate–using a design to accomplish a task not originally intended [Dix, 2007]. Furthermore, we noted that since B2 was a researcher working with gaze data which is a type of spatial data, Gander could also be modified to suit her use. This version of Gander would be similar to STREAM by Hubenschmid et al. [2021] which is an AR+tablet system for exploring trajectory data.

7.5 Discussion and Future Work

7.5.1 Deduction, Customization, and Appropriation

Given the myriad of spatial analysis techniques, it is impossible for Gander to focus on all techniques. Future versions of Gander should aim to be highly customizable, and make it easy for the user to appropriate for other purposes (e.g., teaching, and communicating information to other non-users). Therefore, future versions of Gander should perform the following:

- **Deduction:** If able to, Gander should recommend an appropriate technique to the user.

- **Customization and Appropriation:** Gander should allow the user to program and add functionalities that they require. Alternatively, it can connect with other statistical packages; Gander itself, is already relying on R for statistical computation. Furthermore, some users may want to use Gander for different purposes such as teaching. Therefore, the customization should allow the system to be appropriated as well.

7.5.2 Providing Overview Information

For Gander to meet the requirements of expert users, presentations of overview information must supplement the glyph-based visualization. We found the expert users, particularly after fitting the data, primarily relied on overview information to make decisions. The overview information can come in different forms such as aggregated statistics (e.g., mean, median, mode), or test statistics (e.g. t-statistics). Therefore, future work must investigate proper means of visualizing various types of statistics, and how to incorporate the visualizations into an AR+tablet interface.

Additionally, providing overview information may help to reduce the chance of committing the *atomic fallacy.* According to Keskin [2022], an atomic fallacy is when we conflate a few data to be a general trend. For example, an analyst finds a *few* polluted lakes in a region and declares that *every* lake in the region is polluted. Gander's glyph-based visualization is designed to reduce ecological fallacy. However, as Zhao et al. [2017] argue, humans are good at extracting trends from visualizations, including trends that are not statistically significant. Based on this argument, our glyph-based visualization may instead encourage the user to commit the atomic fallacy. While providing overview information or aggregated statistics alone may encourage the ecological fallacy, it may serve as a good counterbalance. More work is necessary to better understand the relationship between the atomic and ecological fallacies, and how to counterbalance each other using visualizations.

7.5.3 Better Support for Direct Manipulation

While the current Gander interface uses direct manipulation, several participants pointed out that the actions still felt disjointed. For instance, changing the visualization involved too many steps. As such, a future version of Gander should have its Variable Picker and Model Comparer redesigned to ensure a more direct manipulation of the glyphs. To better support more complex statistical techniques and analysis tasks, the future version of Gander can also include a touch-supported visual programming language. An example of this is Microsoft Touch Develop by Ball et al. [2016]. It is a visual programming language designed for touch gestures on tablets. Other AR-based input methods (e.g., voice commands, and hand gestures) may be explored as a supplement to the revised touch-based interaction.

7.5.4 Collaboration

Since Gander can create room-sized visualization, we should explore collaboration. Collaboration allows multiple analysts to co-pilot a statistical analysis. Co-piloting, according to Veldkamp et al. [2014], is where one analyst verifies the results of the other analyst's work. Veldkamp et al. argue that this can lead to better analyses. Furthermore, as we have learned in the study, a geospatial analysis system has multiple stakeholders. In some scenarios, non-primary stakeholders end up being more affected than the primary users themselves. Future work should explore how multiple stakeholders can use Gander at once and benefit from collaboration.

7.6 Conclusion

The study provides us with valuable information for updating the prototype. The low-fidelity version of the updated prototype can be found in Ch. 9. This study also nicely compliments the synoptic and elementary studies which have too much experimental control. Additionally, we also learn more about the potential users of Gander.

Chapter 8

Discussion

In this chapter, we discuss our work, contextualized by the three research objectives outlined in the induction chapter. The three research objectives are:

- **Obj1. Design as research with a vertical slice version of Gander:** This section discusses our findings in the context of developing a vertical slice as a part of "design as research."

- **Obj2. Glyph-based Visualization:** We provide a discussion of glyph-based visualization used in our work. Although the chapters for the synoptic and the elementary studies already contain discussions on this matter, this section serves as a bridge to link those discussions together. We also propose additional enhancements to glyph-based visualization.

- **Obj3. Combining augmented reality headsets and tablet computers:** This section provides a further discussion on combining OST-HWD AR together with a tablet interface, in light of the study results.

While this chapter suggests some new design elements, they are not realized here. Instead, the low-fidelity of the new design ideas are available in the next chapter (Ch. 9).

8.1 Relevant to Obj1. Design as research with a vertical slice of Gander

We have expected the design as research approach to provide some clarity on how a common user can engage with a geospatial system like Gander. The synoptic study, while useful for understanding scanning behaviour, does not reveal how users can accomplish tasks due to the effect size instrument being too insensitive. Furthermore, instead of observing a common pattern among user in the walkthrough demonstration study, we find diversity. Still, these studies reveal some insights about potential users of the system.

8.1.1 Overview Information

Our studies show that Gander's glyph-based visualization may be insufficient. The synoptic study (Ch. 5) shows the participants may have overestimated the effects. While more work is necessary to identify the root cause of overestimation, we think future work should explore providing additional overview information as a way to discourage overestimation. The participants in the synoptic study may have overestimated, because they were noticing trends from the glyphs and concluded that all trends must be real. According to Zhao et al. [2017], this phenomenon is common in visualization.

Meanwhile, several participants of the walkthrough demonstration study (Ch. 6) stated that they wanted to see supplementary overview information, in addition to the glyphs. When a MLR model is created, there should also be a table that describes the MLR coefficients and other relevant information. Therefore, we propose the use of the tablet to help with displaying the statistical information. The tablet, as a device with a higher display resolution than OST-HWDs, maximizes the legibility of the presented information. Future work will involve how the new overview information tables can influence perceptions of glyph-based visualization.

An overview map should also be provided. A participant in the walkthrough demonstration study wanted to be able to see the map as a whole. The ability to zoom out would make the map easier to navigate, and to obtain general information. However, implementing this type of overview requires consideration of aggregation of the glyphs. Designing and implementing glyph aggregation requires thoughtful considerations, because it may impact how the user makes inferences. For example, when a user zooms in, the glyphs can either: maintain the same sizes which increases the space between them, or become larger. These behaviours elicit very different perceptions of the glyphs.

8.1.2 Customization and Appropriation

Future versions of Gander should aim to be highly customizable, and make it easy for the user to appropriate for other purposes. Given the myriad of spatial analysis techniques, it is impossible for Gander to focus on all techniques. Therefore, a user may need to customize Gander. The user may install a plug-in or replace the R script used for fitting models and computing effect sizes. This leads to the topic of mixed reality programming. The participants of the walkthrough demonstrations suggested that Gander should also be used for communication–e.g., to communicate statistical information to non-users, or to teach statistical concepts. The suggestion implied that future versions of Gander should be easily appropriated so that it is useful for other purposes.

8.1.3 Stakeholders

We found that other non-users cannot be ignored. In certain cases, non-users were more impacted by the analysis than the primary users themselves. While the user, the primary stakeholder, generates the models, the secondary stakeholder interprets the models into policies that affect the tertiary stakeholder. If the policies are incorrect, the non-users can be greatly affected–even more than the user. The user may simply be reprimanded while the tertiary stakeholders lose money, time, and resources in order to comply with the policies.

Some primary users may not be advanced enough to use the necessary techniques. Therefore, Gander can help to deduce the appropriate technique for the user. For instance, if the dependent values (DV) are binary (e.g., 0 or 1, "yes" or "no"), Gander can suggest logistic regression to the user.

Since we can expect Gander to have multiple direct and indirect users with varying degrees of experience, collaboration becomes a very important topic. Multiple analysts should work together to correct each other's statistical work as suggested by Veldkamp et al. [2014]. Other stakeholders could be involved. Since Gander provides room-sized visualization, extending it to support collaborative work is relatively straightforward. Multiple works like MARVIS by Langner et al. [2021], and Airbus Tactical Sandbox [Walsh et al., 2023] are already incorporating AR into collaborative decision-making. In addition to collaboration with other human beings, virtual agents can also enhance Gander. Weitz et al. [2019, 2021] explore the use of virtual agents to improve statistical inference.

8.1.4 Other Statistical Techniques: Towards a Full IML System

The current version of Gander can only deal with MLR. We specifically chose the technique, because it allows us to more easily complete the vertical slice. However, in order for Gander to be usable by more users, more techniques must be supported. Some techniques can be added almost right away:

- **Logistic Regression:** For a model with a dependent variable (DV) with values of either zero or one.

- **Generalized Linear Model:** For a model with a DV that, while not normally distributed, is still parametric. For instance, a DV with values bound between an interval can be described using a beta-distribution [Gupta, 2011].

- **MLR with Transformation:** Some IVs or DVs may not have suitable distributions for MLR. However, simple data transformations can make them compatible. For instance, mean-squared displacement (MSD) variables are not normally distributed. However, some of them can be log-transformable [Bailey et al., 2022].

These techniques, while more complex than MLR, still largely resembles MLR. Other regression techniques will require significant redesigns.

Furthermore, we must consider incorporating spatial correlation into the model. The current version of Gander, while involves finding regional differences in the data, does not incorporate spatial autocorrelation into its analysis. There are many types of spatial autocorrelation models; however, the most suitable one for Gander is Geographically Weighted Regression (GWR). According to Comber et al. [2023], GWR involves generating multiple candidate MLR models, and comparing them. The first candidates do not incorporate any spatial information. Later models include regional information as predictors. Eventually, a hierarchical model is created. GWR, unlike other spatial models, is largely exploratory. Because GWR involves multiple rounds of regression and is exploratory in nature, it is suitable to be incorporated into Gander.

Lastly, Gander should seek to expand its capabilities to become an interactive machine learning (IML) system. Since Gander can only perform create ordinary least square (OLS) models, one may not consider it a full IML system. However, as we add more technique, we may match some IML capabilities found in other software like Orange.

8.2 Relevant to Obj2. Glyph-based Visualization

As two of the three studies involve testing glyph visualization techniques, some discussion is warranted. In this section, we discuss the results of the synoptic and the elementary studies together. A bulk of the discussion in this section makes extensive use of semiotics.

8.2.1 Glyph Perception and Semiotics

Visual semiotics play an important role in designing glyph-based visualization. It indicates we can affect the appearance of glyphs (e.g. shapes, colours, position, and more) in order to convey certain information [Borgo et al., 2013, MacEachren et al., 2012]. However, we argue that we can use general semiotics to frame the synoptic and elementary studies.

Semiotics and Semiosis

Up until this point, our use of semiotics is limited to glyph designs. However, semiotics can also be used to analyze other types of visual media. For instance, Dewhirst and Lee [2012] used semiotics to analyze how a cigarette brand in South Korea tried to advertise its product to consumers. Semiotics is also applicable to other senses as well, such as hearing, and smell. However, we shall restrict ourselves to the visual domain here.

A key concept in semiotics is semiosis—i.e. the process of converting one sign to another. In visual semiotics, which is a subset of semiotics, we do not really discuss semiosis. However, the process is implied. For instance, interpreting a glyph's colour into a numerical value in the elementary study is a semiosis, because it involves converting the "colour" sign to the "number" sign. This semiosis itself may require additional semioses to complete the conversion.

It can be difficult to discuss semiosis and semiotics, because many acts that we do are already semiotic in nature. This means, if we casually introduce these concepts without any thoughtful consideration, they do not yield much insight. Instead, they add confusion to the discourse. For instance, when the user fits a statistical model to some data, they are committing an act of semiosis–or the process of converting one sign to another [Martynenko, 2003]. In this case, if we follow the Peircan semiotics, the raw data are a type of sign called the "representamen". The output model is the "object" sign. Meanwhile, the software used to convert the sign is called the "interpretant." In practice, the user would never use "representamen", "interpretant" or "object" to refer to any element of their work. Instead, they would use context-appropriate terms such as "data", "statistical technique", and "test results."

If applying semiotics risks introducing confusion to orderly information, then why should we try to discuss it? As it turns out, Semiotics can help to introduce abstract structure to concepts that have no clear structures at first glance. In our case, using semiotics allows us to frame glyph comprehension even if there are many low-level research gaps (e.g., lack of literature in OST-HWD colour comprehension as pointed by Erickson et al. [2020]). This does not mean we are arguing that the lower-level concepts are totally irrelevant. Rather, the structure allows us to better highlight the potential research gaps that future research should address. Furthermore, using semiotics helps us to establish links with visual semiotics and advances it.

Semiotics has a significant tie to linguistics [Sebeok, 1986]. As such, semiotics shares multiple concepts in linguistics and serves as a more abstract superset of linguistics. In linguistics, there are two major concepts that are relevant to our work: syntactics and semantics. Syntactics discusses how we arrange multiple words together. For instance, "a person pets a cat" and "a cat pets a person" have different meanings–even if they have the same words. Different types of glyph arrangements (e.g., Stacked, Radial) can be considered as different types of syntaxes, and they could impact interpretation. In our elementary study, we noticed that the participants may have made more mistakes with Stacked due to its syntax. Semantics deals with the meanings of words. In a human language, we have words to qualify an amount. In glyph visualization, colour is one of the visual channels that can act as a quantifier.

Syntactics

Syntatics describe the relationships between multiple signs [Zemanek, 1966]. In our work, we explore how the structure of the glyphs can affect their comprehension. In the synoptic study, we focus on comprehension with fields of glyph composites. Meanwhile, in the elementary study, we focus on how a person can understand a composite at a time. In both studies, we discuss how the arrangements of the glyphs affect glyph comprehension.

In our work, we introduce two methods of structuring glyph composites: Stacked, and Radial. Each technique has different syntaxes, and therefore, participants must use different approaches to extract information from the glyphs. In both the synoptic and elementary studies, the participants had legends that they could use to better understand the glyphs' structures. In essence, the legends acted as an interpretant. However, in the elementary study, we found that the user's predisposed bias (e.g. SNARC [Göbel, 2015, Shaki and Fischer, 2018, Aulet et al., 2021]) can act as another interpretant that disrupts the interpretation process. Despite training and the legend, the participants' bias was too strong to be counteracted. Some participants of the synoptic study indicated that they would like the ability to zoom so that they could better separate the glyph composites. The participants of the synoptic study asked for zooming to help separate the glyphs, and to solve the syntax conflict. This means that overdrawing of glyphs can negatively affect the syntaxes of the glyph composites. Another solution to the overdrawing problem is to use different visualizations at different zoom levels. For instance, when zooming out, neighbouring overdrawn glyphs merged into a single shape similar to the splatterplot [Mayorga and Gleicher, 2013]. This destroys the original syntaxes and semantics. However, it encourages the user to interpret the visualization differently.

Semantics

One thing that our work aims to analyze is the semiosis of the glyph colour. We need to know how the user obtains a number value from a glyph's colour, combines it with other values, and eventually transforms it into an insight for the IML system. Our work must also identify various factors, or interpretants, that help or hinder the user. An example of a helpful interpretant is prior experience with Microsoft HoloLens v2.

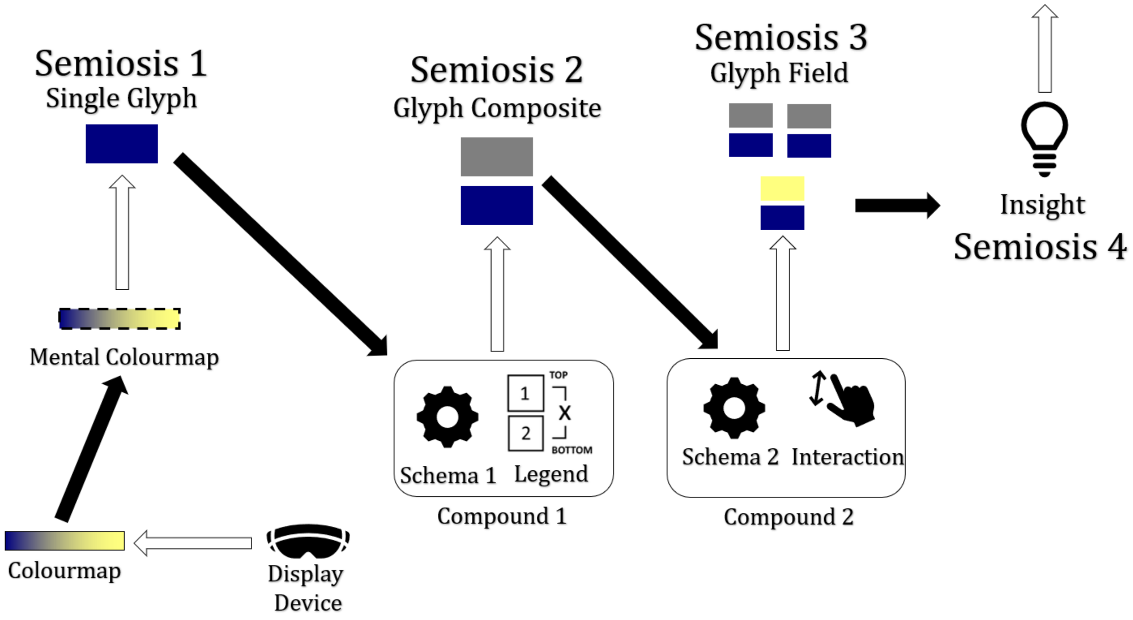


Figure 8.1: A semiosis representing our research work. Hollow arrows represent interpretants affecting a semiosis. Solid arrows represents transformations of representamens into objects.

If a user has some familiarity with the system, they would be more aware of the colour distortion. Therefore, they would be more careful with their interpretation of colours. An unhelpful interpretant could be SNARC which affected how Stacked glyphs were interpreted.

Fig. 8.1 represents a potential semiotic model representing how a user may perform semioses while using Gander. It assumes a bottom-to-top model. The user first interprets the colour of a single glyph, and transforms the number into a schema. How the user interprets the colourmap and applies it is affected by the display device. A schema is a term that we borrow from the cognitive load theory which means a long-term working memory that helps with similar cognitive processes [Paas et al., 2003]. For example, once the user becomes familiar with the colourmap, the subsequent usage may be faster and more accurate. It is important to note that since our study is not longitudinal, it might be unlikely that the participants had developed any real schema.

With a schema, the user then compounds the sign with another sign to create a new interpretant. A compound is a semiotics concept found in linguistics, which refers to multiple words being combined to create a new meaning [Søgaard, 2008]. The meaning can be related, but distinct from the original one. Below are descriptions of

“Legend” and “Interaction”, the two signs found in Fig. 8.1 used for compounding:

Legend. The visual aids for the user—i.e. the colourmap legend in AR, and the composite diagrams in the elementary study.

Interaction. The user’s ability to navigate around the AR setting and its effect on the visualization through the touch gesture on the tablet screen.

The compound becomes an interpretant for the subsequent semioses. Finally, once the user understands the glyph field, they form an insight. This insight becomes an interpretant for understanding the whole ML model.

While we could have expanded the model in Fig. 8.1 to include more semioses, we refrain from doing so. As it turns out, one can easily extend semioses so much that it becomes a series of “unlimited semiosis” [Barr et al., 2004] which is not useful. For this reason, we focus on using semiotics as a tool of convenience for framing our research, and to identify potential research gaps which we have identified as follows:

Colour-Value Judgment. How can Gander assign colours to the glyphs to ensure that the user will be accurate?

Legend Design. How can we design a visual aid that maximizes the effectiveness of Gander?

Interaction. How can interaction help the user with inference?

Correct Insight. How can we measure the user’s understanding of the content presented by Gander, and how to ensure that the user’s understanding is correct? The current measurement method in the synoptic study is ineffective.

The semiotic model (Fig. 8.1) is not necessarily representative of how a user operates. After all, our use of semiotics here is to describe and not to be the description itself. The inability to provide an explanation beyond a surface one is a weakness of semiotics. However, there is an ongoing development to address this issue: cognitive semiotics, a new field introduced by Zlatev [2015]. Unlike the “classical semiotics” which deals with abstract ideas ungrounded by any type of science, cognitive semiotics aims to ground itself with other scientific fields like linguistics, and cognitive science. It also promotes empirical research to validate semiotic explanations. We argue that cognitive semiotics is necessary for creating an effective way to track the user's development of statistical insight while using a system like Gander.

8.2.2 Visualization of Multiplicative Effect

Multiplicative effects are important for fitting modelling. However, as Friedrich [1982], Braumoeller [2004] argue, they can be difficult to understand–even for simple techniques like MLR. We implemented the multiplicative effect visualization in Gander so that the user can better understand them. However, the elementary study shows that the participants tended to choose zero for the multiplicative effect. As it turns out, when multiplying unit interval numbers together, their values tend to be close to zero. We realize that individual values, and their multiplication are different. The former is a length, and the latter is a hypervolume. Therefore, they should be interpreted differently. Different colourmaps or turning a hypervolume into a length may be necessary.

8.2.3 Likelihood Information

The current version of Gander uses E_L to show the goodness of the models. We chose this value for two reasons. First, it is normalized to be between zero and one; thereby, keeping the pre-fit and the post-fit stages consistent. Secondly, it allows for the shader-based effect to show differences between the models. We believe that keeping both stages consistent can aid the user in understanding the model. However, the walkthrough demonstration results suggest that keeping both stages consistent is not important, because the user sees pre-fit and post-fit stages as being highly distinct from each other.

Therefore, we adopt a different measure for comparing models. In the future version of Gander, we plan to use the Akaike Information Criterion (AIC) or other alternatives like the Bayesian Information Criterion. Like E_L, AIC still relies on likelihood [Snipes and Taylor, 2014]. However, unlike E_L, AIC allows for multiple models to be compared at once, thus reducing the repetitions. Furthermore, the user can rely on ranking AIC which makes the process more intuitive [Snipes and Taylor, 2014]. AIC also penalizes non-parsimonius models [Snipes and Taylor, 2014]. The next chapter (Ch. 9) contains information on how we redesign the post-fit visualization to support

AIC.

While the new AIC-based method is very different from the E_L-based one, it still preserves some of the post-fit features. For instance, it still allows for local goodness-of-fit values to be displayed.

8.2.4 Colourmap Design

This work reaffirms a well-known wisdom: colourmaps matter for the user to comprehend the visualization that they are working with [Harrower and Brewer, 2003, Crameri et al., 2020]. However, it also challenges the notion that the rainbow colourmaps are always inappropriate for visualization because they are not as “sortable” as the divergent colourmaps [Crameri et al., 2020, Stoelzle and Stein, 2021]. Instead, we may need to reconsider rainbow or multi-hue colourmaps. After all, Gołębiowska and Çöltekin [2022] argue that the nature of the tasks determines the appropriateness of the rainbow colourmaps; Quinan et al. [2017], Reda and Szafir [2021] found that the rainbow colourmaps can help the user to better dissect a colourmap into a smaller quasi-maps. A multi-hue colourmap may have improved the participants' experience in the synoptic study, and made the answers to the value-judgement task in the elementary study less binary.

Our work further also emphasizes the impact of OST-HWDs on colourmap perception. Although displayed colours varied among multiple types of display devices [Harrower and Brewer, 2003], we found that OST-HWDs were extremely difficult to adjust for. First, the contemporary OST-HWDs tie luminance together with opacity [Itoh et al., 2021]. As such, dark objects appear transparent. Luminance is extremely important for designing a colourmap, particularly the one for the user with CVD [Crameri et al., 2020]. Therefore, if we were to have an accessible colourmap for OST-HWDs, it is imperative that the devices must be able to separate luminance from opacity. According to Itoh et al. [2021], this necessitates the development of OST-HWDs with subtractive screens–i.e. screens that are able to remove light from the physical world. Another issue is that the OST-HWD hardware can distort the colour observed by the participants [Itoh et al., 2021]. Based on the observation in the synoptic study, we found that the participants noticed colour changes by rotating their heads. Multiple colour-adjustment algorithms [Itoh and Klinker, 2015, Kim et al., 2019] exist; however, they remain experimental at this point. Being able to accurately display colour is key to immersive analytics as an inaccurate display of colour can break the sense of immersion. Instead of peacefully viewing the virtual content, a varying and inaccurate colour display device can force the user to doubt what they are observing and constantly readjust themselves.

8.2.5 Atomic Fallacy v. Ecological Fallacy

The glyph-based visualization of Gander, as outlined in Ch. 4, has a fundamental flaw. Since it does not aggregate data in any way, the user can easily make many comparisons. Based on the comparison, the user can then make a conclusion that specific trends exist even if the trends are spurious. For instance, in a region with too few glyph composites to make a concrete inference, the user can conflate the information of these glyphs as a real trend and apply it to the whole region. This error is also called the *atomic fallacy* [Keskin, 2022].

Aggregation (i.e. computing and presenting average) can help data exploration safer. Averages can "blur" the various glyphs to make trends less noticeable to the user. We can also use them to compute p-value which is useful for designing an interface to discourage false discoveries [Zhao et al., 2017]. However, this also introduces an opposite issue. According to Salkeld and Antolin [2020], solely relying on aggregate information can lead to a type of logical fallacy called *ecological fallacy*–when we interpolates the average of a model to individual data point.

We think that by providing both individual and aggregated data, the user can balance between committing an ecological fallacy, and making false discoveries. Providing an overview information window alongside the glyph, as suggested by some participants in the walkthrough demonstration study, can alleviate the issue of discovering false trends. Meanwhile, the glyphs tampers down the ecological fallacy.

8.2.6 Interaction for Better Inference

Not only interaction is a key element in an immersive analytics system, but it could also play a role in resolving visualization issues. A study by Duncan et al. [2021] states that adding interaction can resolve the issue of cartograms' sizes biasing the user's judgments when performing synoptic tasks. The inflated results of effect size questionnaires in our synoptic study hint that the participants may have been making false discoveries. Therefore, adding interaction that can alert the user may be helpful in this regard. Furthermore, Mayorga and Gleicher [2013] argue that adding interaction can help to resolve overdrawing issues. In our studies, we carefully chose maps to minimize the overdrawing of the glyphs. However, in practice, the user cannot choose the data that they deal with. If the distributions of glyphs lead to overdrawing, the user cannot discard the map. Instead, they must find a way to resolve the issue. Therefore, a possible future research direction involves studying multiple types of interaction to improve inference.

Furthermore, we found that the glyph design could also influence how the user interacts with the system; Polyline causes the user to examine glyphs more closely, and Radial causes the user to engage more with visual scanning. Meanwhile, Stacked influences the user to rotate their head more in order to compose and decompose glyphs. Different navigational behaviours lead to different styles of glyph field explorations. This subsequently leads to different ways the user interacts with the overall interface. Future work should explore how visualization and interaction affect each other.

8.3 Relevant to Obj3. Combining augmented reality and tablet display and inputs

8.3.1 Focus+Context

At first glance, Gander seems to follow the focus+context (F+C) paradigm. The tablet, with a high display resolution, serves as the focus resolution. Meanwhile, the AR interface, which has a poorer display resolution, serves as the context resolution. However, to be a true F+C system, the user must be constantly focusing on the tablet, and only rely on the AR for context information. We did not observe this in our synoptic study. Instead, the user mostly relied on the tablet as an input device. To make Gander into a true F+C system, we must force the user to focus on the tablet. There are multiple ways to achieve this. For instance, we can blur the glyphs outside the tablet's boundary to ensure that the user must focus on the tablet in order to examine the glyphs. Alternatively, designs could be used to incentivize focusing on the tablet. For example, due to the parallax effect, Stacked glyphs are automatically decomposed when further away from the user. If the user is only allowed to move beyond the tablet, then they must focus on the tablet to get the blended values of Stacked glyphs.

Alternatives to Focus+context

There are other alternatives to F+C, such as zooming and overview+detail. Zooming is a common type of interaction available in geospatial software, and it was often requested by the participants in the synoptic and walkthrough demonstration studies. Furthermore, since zooming allows the user to easily view and compare glyphs, it can inflate the false discovery rate. Overview+detail (O+D) is another candidate. In this case, the user gets to see overview information in a smaller screen (or window), and detailed information in a larger one [Cockburn et al., 2009]. If we modify Gander to support O+D, the tablet could provide overview information by showing a mini-map of the whole workspace, and other overview statistical information. Meanwhile, the AR displays the glyphs, the map, and other minute details.

8.3.2 Glyph-field Scanning and Navigation

We found that variations in glyph-based visualization can influence how the user interacts with AR+tablet glyph fields. Glyphs that rely on colour channels tend to encourage more visual scanning. The user tends to pan the map less and prefers glancing at the map. Meanwhile, glyphs that use shapes to convey information tend to encourage closer examinations. As it turns out, shapes tend to be more affected by distance than colour. As a result, the user tends to pan the map more to bring the glyphs closer to themselves.

Knowing that we can use different visual channels to manipulate navigation behaviours has a significant implication on interface design. If a designer wishes for the user to visually scan a glyph field, they can use colour glyphs. If the designer wishes for the user to pan the field, shape-based glyphs can be used. Future work should explore the impact of other visual channels and glyph navigation behaviours. Furthermore, the work should study how allowing the user to conveniently switch between multiple types of glyphs can help to navigate and understand the glyph field.

8.3.3 Multiple Input Methods

We note that we never used the HoloLens v2's inputting method for our studies; the user can only manipulate Gander from the tablet. This is partially motivated by Feiner and Shamash [1991], Soares et al. [2021] who found AR inputs to be imprecise. In the future, we should explore comparing and combining both input methods in order to maximize the potential of a multi-device hybrid user interface. Another aspect of input that we should consider is direct manipulation. Direct manipulation exists to a degree within Gander. However, the participants of the walkthrough demonstration study still found updating AR glyph fields to be cumbersome. Future versions of Gander should explore more instantaneous updating of AR glyph fields through both AR-based and tablet-based input methods.

8.3.4 Mobility

Both untethered OST-HWDs and tablets are mobile devices. By this logic, Gander could also be categorized as mobile software since it relies solely on mobile devices. However, it has some limitations. Although Gander is mobile enough for us to set up in multiple places for our studies, in its current form, it is not sufficiently mobile to be deployed in an arbitrary environment.

There are several major hurdles that we must overcome before we can create a truly mobile version of Gander. First, OST-HWDs can be impacted by the light of the physical world. A bright sunny day prevents the devices from properly rendering virtual content. Secondly, a good network connection is necessary. Whitlock et al. [2020] note that setting up a good mobile network infrastructure, particularly for outdoor environments, can be extremely difficult. Thirdly, Gander is not fully spatially aware. After the tablet's position is synchronized with the OST-HWD, they never relay their positions again. While the HoloLens v2 is extremely capable in terms of tracking itself in the physical world, it cannot track other devices in real-time. Solutions exist (e.g. QR-code tracking); however, they introduce their own sets of limitations.

Chapter 9

Proposed Changes to Gander

Based on the study results, we propose multiple modifications to the current version of Gander. Unlike Ch. 4 which describes a high-fidelity vertical slice prototype, this chapter contains a set of proposed design changes to expand the vertical slice–making it more functional. We used Miro [Miro, 2023] to create the low-fidelity designs. The updates that we propose aim to combine the granular glyph-based visualization with other visualization techniques that use aggregated information in order to provide an optimized user experience. Furthermore, interactivity is increased to not only make the interface more pleasant to use, but also to help with statistical inference.

Unlike the current design which the AR content is always mounted on top of the tablet, the system adapts to the tablet's position. For instance, the user can unmount the AR content from the tablet in order to use Gander in the overview+detail (O+D) paradigm. In this way, the tablet provides a mini-map while the user can look for details using the AR-based visualization. Alternatively, mounting the AR content on top of the tablet turns Gander into a focus+context (F+C) interface. The tablet now provides the details by showing glyphs that are more suitable for closer examination such as Polyline.

Fig. 9.1 shows an evolution of Gander's tablet interface when being used in an O+D mode. We envision the user holding the tablet with the AR content unmounted from it. Instead of scrolling the screen to move the map, the four-way arrow at the bottom-left of the screen can be to pan the map, in a similar manner to a pointing stick on a ThinkPad. We propose the use of a pointing stick for moving the AR content because we found long swipe gestures to be uncomfortable during the pilot study for the synoptic study.

Instead of relying on the Variable Picker to slowly manipulate the glyphs, the user can use short swipe gestures to rearrange the order of the glyphs in "Layers." Each layer has toggles that allow the user to temporarily hide variables without removing them. We implemented this functionality because of the elementary study which shows that static visual aids may not be accurate. Adding interactivity can help to

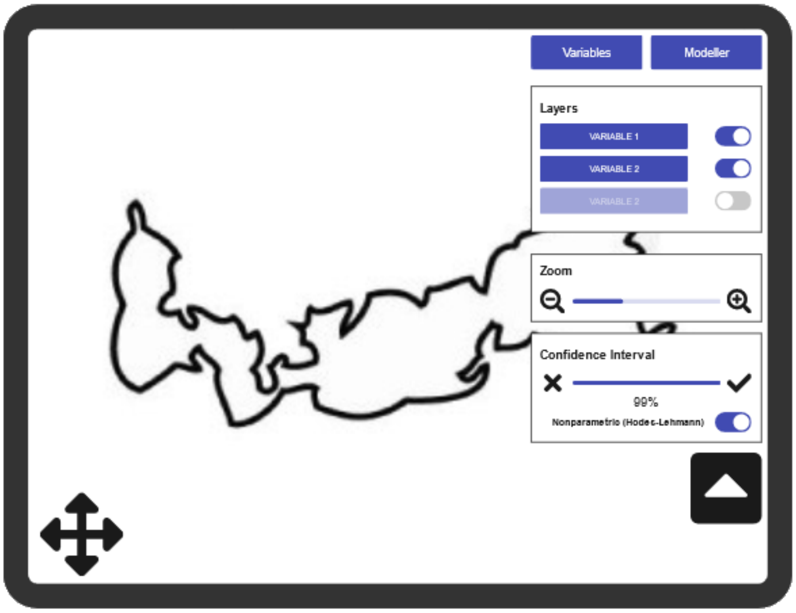


Figure 9.1: The new design of Gander's tablet interface for O+D display. The map outline is from [Immigration, Refugee, and Citizenship Canada, 2020].

improve the comprehension of a visualization [Duncan et al., 2021]. Furthermore, the participants of the walkthrough demonstration study (Chapter 7) pointed out that the Variable Picker was too cumbersome to use.

Additionally, there are two extra panels for controlling AR plots (Sec. 9.1.2). "Zoom" scales the AR content as well as the AR plots. Furthermore, there is a panel for indicating the confidence interval. While the default confidence interval is 95%, the user can adjust the interval. If the user wishes to compute the confidence interval using a nonparametric method, Hodges-Lehmann estimates (see Hodges Jr. and Lehmann [1963]) can be used instead. Bootstrapping is an alternative to the Hodges-Lehmann estimate; however, it might require too much computational power from the hardware. These interface elements are inspired by our previous work on "average-based selection" in Ch. 3, and feedback from the participants in Ch. 7.

9.1 Aggregation and Safe Exploration

Computing aggregated information, as indicated by the participants of the walkthrough demonstration study, is important. While glyph fields are valuable, they must be supplemented. Before we can create aggregates, we must develop a way to

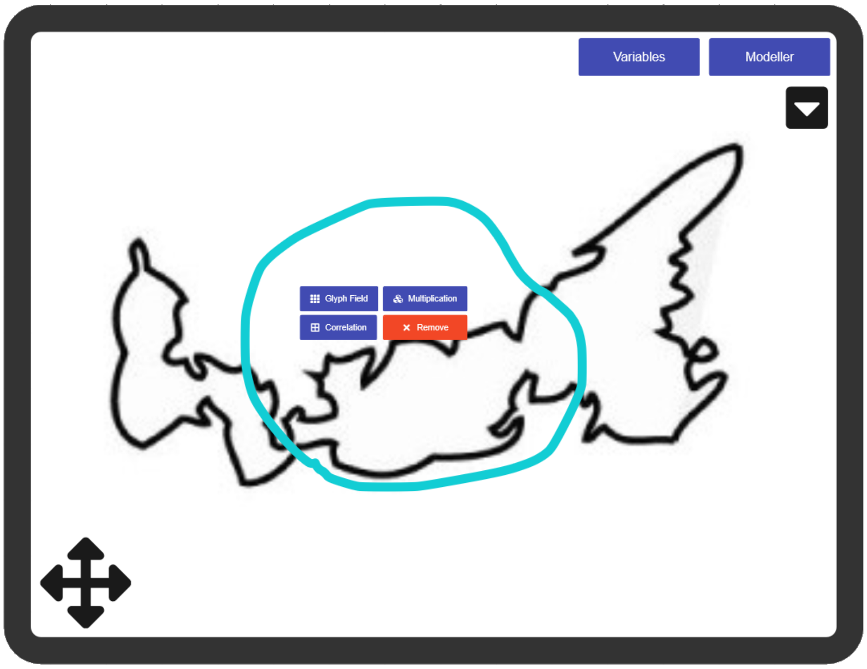
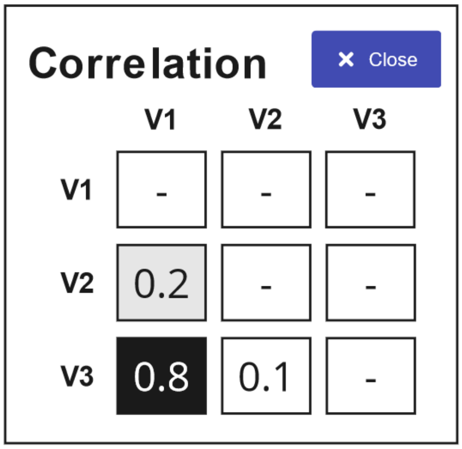


Figure 9.2: The tablet interface of Gander with a part of Prince Edward Island selected. The four buttons show up to allow the user to examine the cut. "Glyph Field" allows the user to toggle between the glyph and the pancake plot display (see Sec. 9.1.2. "Multiplication" shows an overview window for multiplicative information. "Correlation" shows the correlation matrix between the selected variables. "Remove" eliminates the cut. The side menu is hidden. We use an artwork from [Immigration, Refugee, and Citizenship Canada, 2020].

subdivide the map. As mentioned in the walkthrough study chapter, this process can either be automated or the user can manually segment the map. Since a tablet is a more accurate input device than OST-HWDs [Soares et al., 2021, Feiner and Shamash, 1991], we propose allowing the user to make cuts using touch gestures on the tablet like in Fig. 9.2. Assuming that Gander is in the O+D mode instead of the F+C mode, the user can use the tablet to select the cut that they want to review. Then, they can select a dialog box that presents the summary information. Section 9.1.1 has information on these dialog boxes—including how they can appear during the pre-fit and the post-fit stage. Meanwhile, Section 9.1.2 describes a new type of visualization for Gander: the pancake plot. The pancake plot is a 3D cartogram for the AR interface and it serves as an aggregated alternative to Stacked.

9.1.1 Complex Information Windows

In the walkthrough demonstration study, some participants expressed a desire to have a statistical table or charts available to them. Furthermore, the synoptic study shows that the glyph-based visualization may encourage false discoveries; therefore, having overview information available can serve as a guardrail against this. We propose that in the pre-fit stage, there should be an overview window for: correlation, and multiplicativity. When the user selects a cut, they open a window. Fig. 9.3 shows a potential for a window that presents an overview correlation for selected variables in the cut.



	V1	V2	V3
V1	-	-	-
V2	0.2	-	-
V3	0.8	0.1	-

Figure 9.3: Correlation matrix window.

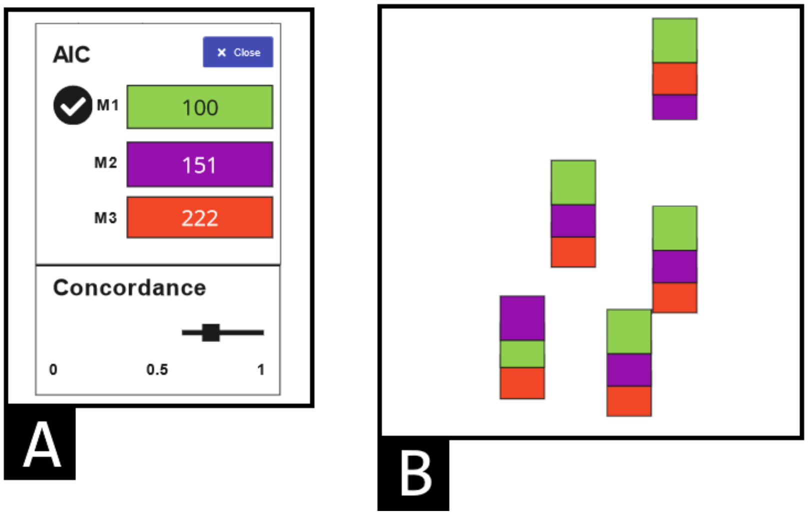


Figure 9.4: **A:** An overview window for the new AIC-based visualization. **B:** The glyphs with the AIC information. The height indicates the rank of the AIC. The colour indicates the model that the glyph is associated with.

In the post-fit stage, we can have a window representing all average goodness-of-fit information in a cut. Alongside the window, the user can also access the tables that describe the model's other statistics such as coefficients. Instead of using E_L, future versions of Gander will use AIC (or similar measures like Bayesian Information Criterion) to generate the glyphs. AIC allows for simultaneous comparisons of all techniques. In the new AIC-based visualization system, we compare multiple models at once. The height of a glyph in a composite indicates ranking. A higher glyph in a composite means the model has a better rank (Fig. 9.4-B). Glyph colours, derived from a categorical colourmap, indicate the model. Although using shapes may be better, we must be mindful of the AR performance. OST-HWDs, like Microsoft HoloLens v2, work best with minimal numbers of vertices. These devices may struggle to draw glyph fields full of shapes other than triangles and/or squares. For overview post-fit information, a future version of Gander can provide average AIC statistics for the selected area. Furthermore, that version of Gander can display Kendall's W. According to Marozzi [2014], Kendall's W is a value between zero and one that the agreement of "judges" on the ranking of specific numbers of "items." One represents complete agreement, while zero represents full discordance. In the context of Gander, a "judge" is a spatial point and the "items" are the models being compared. Since we can easily derive a χ^2-statistic from a Kendall's W, it is easy to generate a confidence interval for Kendall's W as well. Fig. 9.4-A shows a prototype of how the average AICs and Kendall's W could be presented to the user.

9.1.2 Pancake Plot

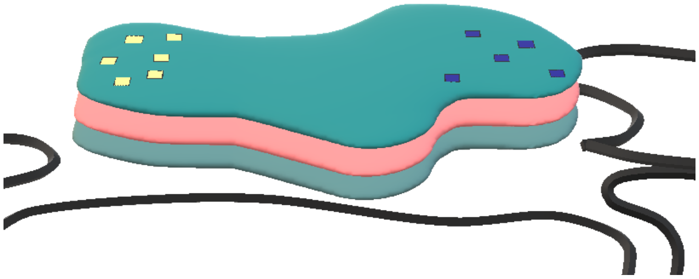


Figure 9.5: Pancake plot. In this figure, the pancake parts use the colourmap from the elementary study and the chips use the colourmap from the synoptic study. The black lines represent other cuts in AR without any pancake plot. In this prototype, the pancakes use Saga colourmap, and the chips use Ukraine colourmap.

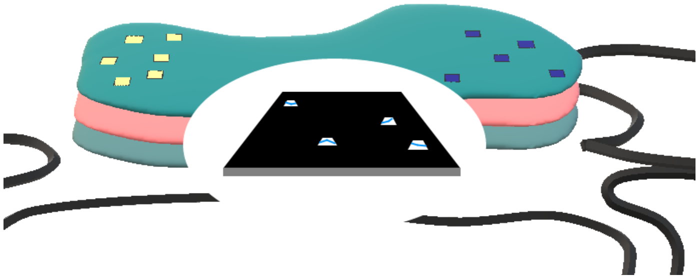


Figure 9.6: Pancake plot with the tablet in the F+C mode. The plot is cut out so that the AR content does not overlap the tablet's content. Polyline glyphs are displayed here because we assume the user wants to glean as much information as possible.

A pancake plot is essentially a 3D cartogram alternative to the glyph-based visualization. It is displayed when the user is trying to see aggregated information. Instead of representing a single point, it occupies the whole selected cut in multiple layers. Glyphs, adorning the pancakes, are used to indicate extreme values and outliers. We call these glyphs "chips." If the user decides to switch to the F+C mode by placing the tablet in the vicinity of the pancakes, a cut-out is created, allowing the user to glyphs representing the individual data used to create the aggregates (Fig. 9.6). The user can also hide the pancake plot in order to see the original glyphs.

One may note that the plot bears some similarity with Mayorga and Gleicher [2013]'s splatterplot. However, unlike the splatterplot, a pancake plot is based on cartograms instead of density plots. Furthermore, the pancake plot is a 3D visualization that requires user interaction to resolve the layers. The interaction should also reduce the size-based bias confound in the cartogram [Duncan et al., 2021].

Pre-fit Stage

In the pre-fit stage, the colour of each layer represents an average. If the Hodges-Lehmann estimate setting is enabled, a pseudo-median is used. Otherwise, the colour represents the mean. Glyphs, representing values that do not fall between the average's confidence interval, adorn the pancakes. We call these glyphs, "chips." These represent extreme values, and the user should analyze them further. The pancake and the chips use different colourmaps to distinguish themselves. The user can swipe on the variable list to change the order. Fig. 9.5 shows a low-fidelity prototype of the pancake plot. We can see that for this particular pancake plot, the left and the right sides contain extreme values.

Although Fig. 9.5 is using the Saga, and Ukraine colourmaps, we are not suggesting that these colourmaps are ideal. In fact, our studies suggest that a multi-hue rainbow colourmap may be more appropriate because it might discourage the user from making binary colour-value judgments. Before implementing a higher fidelity of the pancake plot, we suggest empirically evaluating additional colourmaps and replacing Saga and Ukraine with the best colourmaps determined by the evaluation.

To visualize the multiplicative effect, we may need to devise a novel approach. While the elementary study (Chapter 6) suggests using a separate colourmap, or taking the n-th root of the overlapping value, the effectiveness of the methods must be evaluated before we can implement it. Whether we decide to implement the new colourmap or not, a complex information window should also be used.

Post-fit Stage

In the post-fit stage, we first compute the median AIC values of the models in the selected cut. Then, Gander ranks the AICs and assigns the ranks to the corresponding pancakes' height. Each model is assigned a colour from a categorical colourmap; therefore, the pancake's colour is based on the model.

To create a chip in the post-fit stage, we follow a different set of assignment procedure as outlined in Fig. 9.7. For each model m and for each data point x_i, we first compute an AIC value for the model which we call a. Then, we check if a is

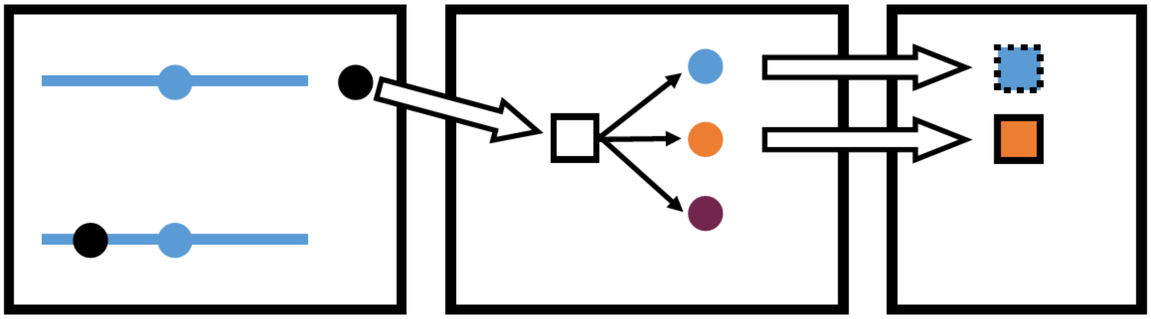


Figure 9.7: An outline of how a chip could be assigned its colour. **LEFT:** a is compared against m's confidence interval. If a is within the confidence interval, a chip is assigned. Otherwise, there is no assignment and the procedure ends. **MIDDLE:** Once a chip is assigned, a is further compared against the average AIC values of all models. If a is still closet to m's average, then it receives m's colour. Otherwise, it is assigned the colour of the other model. **RIGHT:** If the chip receives m's colour, the shape is slightly distorted. Otherwise, there is no change.

outside the confidence interval of m's average AICs. A chip then is only created if the value is outside the confidence interval. Otherwise, we proceed to the next data point and/or the next model. Afterwards, we assign the chip the colour of the model (does not have to be m) whose average AIC value is closest to a. If m is selected, we assign the chip the same colour as m's colour. However, we add a thick border to the glyph to denote an outlier. Since the colourmap is now categorical, we remove the shader-based subtractive blending used in the original version of Gander.

Since there is no parametric method to construct a confidence interval for AIC, we will need to use a nonparametric one like the Hodges-Lehmann method.

9.2 Visual Language for Data Selection and Modelling

The tablet interface of Gander contains several dialog boxes. We argue that some dialogue boxes limit user interaction. Particularly, the user cannot perform any data processing task in the data selection stage, and they are limited to MLR when fitting the data. Therefore, we propose replacing the data selection screen with a visual query language, and the modeller with an interface similar to Orange [Demšar et al., 2013]. The prototypes provided here (Fig. 9.8 and Fig. 9.9) can be implemented either on the tablet or in AR.

9.2.1 Visual Query Language for Data Selection

The current version of Gander only allows the user to select data. They cannot modify the data nor filter them. They also cannot join data sets together. By introducing a visual query language (similar to the one found in Ch. 3), the user can use interactive widgets to select the data that are relevant to them. For instance, they can remove data attributes (i.e. columns) that are not relevant to them before the pre-fit stage. They can also remove rows of data that do not meet specific conditions, allowing the user to create models that are conditioned on certain properties. Fig. 9.8 shows an example of a visual query language for selecting ship data from a geospatial database, filtering out small ships, and presenting the data in the next stage.

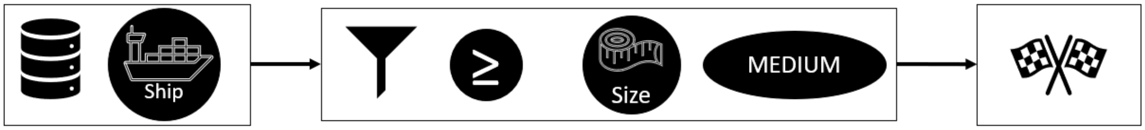


Figure 9.8: A prototype of the data selection widget based on the visual query language.

9.2.2 Orange-Inspired Modeller

Orange is a graphical desktop-based data mining tool Demšar et al. [2013]. The user creates a script by placing nodes and connecting the nodes with lines. Each node represents a ML task while each line represents a flow of information. Orange essentially turns data analysis into a network graph. In the future, Gander can expand the Modeller dialogue to be more graphical, and more like Orange.

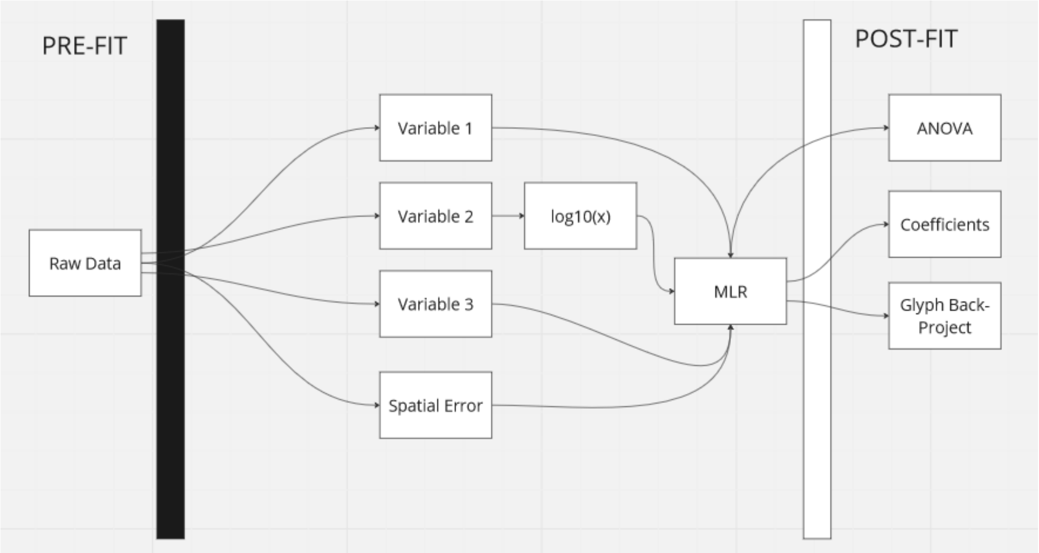


Figure 9.9: A new Modeller prototype inspired by Orange. The user relies on a direct manipulation (drag-and-drop) to move nodes onto the work area. They then draw lines between the nodes. In this figure, we are creating a three-variable MLR model with spatial lag. The second variable is log-transformed. In the post-fit stage, in addition to the glyph-based visualization, the user is also asking for an ANOVA table and a coefficient table.

Fig. 9.9 shows a low-fidelity prototype of the new Modeller. The Modeller shows the gulf between the pre-fit and the post-fit stage. The user develops a network to cross the gulf. Unlike the current version of the Modeller, the user has more choices in what they want to do. For instance, they can add nodes for spatial analysis and data transformation. Since not all users analyze their models in detail, the post-fit glyph-based visualization is now optional. However, if the user wishes to do so, they can add a node called “Glyph Back-Project.” Novice users can ask Gander to prepopulate the field. On the other hand, more advanced users who cannot find suitable nodes can program their own.

9.3 Collaboration, Avatar and Artificial Colleagues

The walkthrough demonstration study shows that geospatial analysts do not work in a vacuum. While Gander targets advanced users, the other stakeholders still play an important role. Therefore, they should not be excluded. Future work should explore multiple users interacting with Gander at once. We identify the following types of collaborations: co-piloting, and multiple types of stakeholders. Co-piloting is a practice where multiple people with the same statistical background scrutinize each other’s statistical analysis [Nuijten et al., 2016]. This practice is encouraged in the statistical analysis of psychological study data to minimize inferential mistakes [Nuijten et al., 2016]. Researchers (e.g., [Du et al., 2018, Langner et al., 2021]) have been exploring the use of mixed reality to assist multiple stakeholders in communicating with each other.

If the co-pilots or stakeholders are not available in person or as avatars, we should explore the use of a mixed reality agent (MiRA). A MiRA is a virtual 3D character that assists the user in the mixed reality environment [Holz et al., 2011]. Its main task is to provide human-like interaction to the user of the system. A benefit of MiRA is that since it most likely to be AI-based, it can make better recommendations for statistical techniques than a human [Weitz et al., 2019, 2021]. However, this does not mean it is perfect, because all statistical techniques have certain amounts of errors (e.g. Type I Error). Therefore, the MiRA must have a mechanism to inform the user about potential errors and uncertainties.

Chapter 10

Conclusion

AR and OST-WHDs afford us new opportunities for geospatial analysis. Such technologies enable analysts to become highly mobile. Some researchers [Whitlock et al., 2020] argue that these devices enable geospatial analysts to work in the field, and in-situ–thus improving their data analysis process. However, due to their novelty, there is much that we do not understand. For example, we still do not understand how the user navigates an immersive visualization using such an interface. To address the research gap, we developed a prototype called Gander. Gander was developed as a vertical slice in order to showcase the completion of the main tasks–thus, making it easier for prospective users to understand how they incorporate such a system into their workflow.

Before we could design a system like Gander, we must first perform exploratory work in immersive analytics to identify the areas of focus. The exploratory work was necessary to pick the features for the verticle slice. Prior to the design of Gander, we conducted research in visual analytics, and immersive analytics. We prototyped an interactive visual query system, and worked on improving user experience within an immersive analytics system. During this phase, we successfully conducted and published studies on visual cueing techniques for immersive analytics [Hu et al., 2021].

Once we identified a multi-device hybrid user interface for geospatial analysis as a research topic, we then engaged in design as research to make sure that the steps in Gander track with existing statistical procedures. The design process of Gander follows the vertical slicing principle. Gander is not only a high-fidelity prototype. It has a sufficient amount of functionality that a real user can perform a geospaital task from the beginning until the end [Ratner and Harvey, 2011]. Having a functional prototype also allows us to showcase various aspects to our participants, and allows us to file a patent [Hu et al., 2022] with the support of Global Artificial Intelligence Accelerator at Ericsson. Furthermore, we explored the use of glyph-based visualization with advanced blending techniques to support detailed visualizations of data, and goodness-of-fit of statistical models.

Once the design process was completed, we conducted multiple studies with human participants to validate the design. Multiple studies, conducted as parts of this work, focused on several aspects of Gander. The synoptic study (Ch. 5) focuses on how the user can glean information through glyph fields. We found different glyph-based visualization techniques induce different navigational behaviours. This has a significant implication for designing glyph fields. We also found that using multiple-choice questions to evaluate the user's understanding of a glyph field is not effective. While this is successfully deployed in multiple information visualization works (e.g. Lee et al. [2017], Peña-Araya et al. [2020]), we must consider a new approach for immersive analytics.

The elementary study (Ch. 6) is about understanding how the user gleans information from the glyphs at the level of a composite. Unlike the previous study, the participant could only see four glyphs at most. The study investigated the impact of using the parallax effect to compose and decompose Stacked glyphs. We found the parallax effect used by Stacked made the user faster at the colour-value judgment task used in the study. However, the technique was less accurate because the user had a harder time comprehending the structure. More glyphs inside a composite made a technique less effective. Furthermore, using intersecting areas of the glyphs to convey secondary information is not a viable strategy. As more unit interval values were multiplied, the numbers became closer to zero. Therefore, the blended values must be adjusted and/or presented differently. Lastly, our work suggests a divergent colourmap may not be viable, because the participants tended to make binary decisions based on the two extremes.

The walkthrough demonstration study (Ch. 7) involves presenting and discussing the prototype to potential users. Despite the small sample size, we discovered that the potential users of Gander are a diverse group of people. Each geospatial analyst relied on different sets of procedures based on multiple factors–from the type of data to the decision-makers who rely on the results of their analyses. As such, future versions of Gander should support easy customization, and collaborative work. In terms of the interface itself, the participants indicated that the tablet interface should be more interactive. Additionally, overview information (e.g., statistic tables) should be shown in addition to glyph-based visualization as reviewing overview information is standard in geospatial analysis.

Using the results of the studies, we propose multiple changes to Gander to expand the vertical slice. Several low-fidelity prototypes are provided to indicate the changes. The new version of Gander allows for both aggregated information and granular information to be displayed. The tablet interface is also more interactive to improve the user experience.

Advancements in geospatial analysis technologies are extremely important, because geospatial analysis is often deployed in decision-making processes that have visible real-life impacts. Our work represents advancements in multiple areas. The design of Gander advances GIS through the introduction of a mobile in-situ large-area display for geospatial analysis. Furthermore, the design incorporates the use of the parallax effect–something that is impossible without an OST-HWD. The synoptic study represents an advancement in hybrid user interfaces by showing that we can use glyph design to control the user's glyph field navigation behaviours. The elementary study advances immersive analytics by analyzing how a user comprehends glyphs that can be composed and decomposed using the parallax effect. The results of the walkthrough demonstration study provide some feedback on Gander itself. Lastly, we propose new changes to the design based on the results of the studies.

10.1 Positionality Statement

I am a researcher with an interdisciplinary background. I equally value science and humanities. During my Bachelor's degree program, I studied computer science (major), cognitive science (major), and French (minor). I also performed research work in cognitive science and computer science. Additionally, thanks to my French program, I learned about semiotics–hence its extensive use in this thesis.

During my Master's degree program, I worked on computer-assisted language learning to align with my interest in language learning. As I was a biracial international student, learning language was something that I was extremely familiar with. When working to develop a video player for language learning, I became interested in research methodologies, and establishing scientific validity. Unfortunately, due to time constraints posed by the length of the program, I was unable to evaluate the prototype to the scientific rigour that I wished for.

In my PhD program, I decided to embark on working with MR. I was excited by the new technologies and their endless possibilities. My original plan was to continue on working an educational project with immersive technologies. Instead of simply focusing on language learning, I expanded my area to general education in the hope that I could find a sub-field of education that could be more easily validated. However, my interest in statistics and scientific methods during my Master's degree kept growing until it supplanted my original goal completely.

Fortunately, my interests in statistics aligned well with my laboratory's priorities. As it turned out, the laboratory had been actively engaging with organizations that relied on data science such as Lockheed-Martin, and Ericsson. Therefore, Gander was an easy project to propose.

My analysis methods reflect my interest in statistics. I am not simply content with the standard conventional tests that I was taught with. Instead, I researched how the various statistical methods can be incorporated. I also insisted on finding out about the origins of the methods to better understand their logic and justification. For instance, I researched the history of the t-test, the Wilcoxon signed-rank test, and more. While this had occasionally led me to deviate from the straight path to completing the programs, the additional research did help me to better understand the potential users of statistical software like Gander.

Although I no longer directly work in the field of computer-assisted language learning and cognitive science, I still have not completely abandoned the roots. Therefore, elements of these two fields manifest within the work. An example of this is when I used SNARC to explain user behaviour in the elementary study. Furthermore, as a former cognitive science student, I am very much interested in how a person can glean information at the smallest scale and convert what they learn into insight. I was somewhat disappointed that I was not able to provide a mechanistic explanation of how a user can transform the glyph information into a statistical insight. This will be an opportunity for future researchers to address.

I believe that for MR to be usable to everyone, and for the goals of the metaverse to be achieved, MR technologies must be inclusive. Lack of inclusivity means the technologies will be out of reach for many. This thesis does not reflect my desire for diversity. The synoptic study was conducted with a predominantly male panel of participants, and the elementary study excluded those with colour-vision deficiency (CVD). I hope that future research will be able to address this.

Bibliography

A. Turner, v-chmccl, and V. Tieto. Holographic Rendering overview. *Mixed Reality documentation*, Jan 2022. URL https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/rendering-overview.

N. Adnan, M. Ahmad, and R. Adnan. A Comparative Study On Some Methods For Handling Multicollinearity Problems. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 22:109—119, December 2006. doi: 10.11113/matematika.v22.n.179.

P. J. A. L. Almeida, M. V. Vieira, M. Kajin, G. Forero-Medina, and R. Cerqueira. Indices of movement behaviour: conceptual background, effects of scale and location errors. *Zoologia (Curitiba)*, 27(5), 2010. doi: 10.1590/S1984-46702010000500002.

M. J. Anderson. Permutational multivariate analysis of variance (permanova). In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, 2017. ISBN 9781118445112. doi: 10.1002/9781118445112.stat07841.

N. Andrienko and G. Andrienko. Tasks. In *Exploratory Analysis of Spatial and Temporal Data : A Systematic Approach*. Springer Berlin/Heidelberg, Berlin, Germany, 2005. ISBN 9783540311904.

ArcGIS. Esri color ramps, 2022. URL https://developers.arcgis.com/javascript/latest/visualization/symbols-color-ramps/esri-color-ramps/.

L. S. Aulet, S. R. Yousif, and S. F. Lourenco. Spatial–numerical associations from a novel paradigm support the mental number line account. *Quarterly Journal of Experimental Psychology*, 74(10):1829–1840, 2021. doi: 10.1177/17470218211008733.

M. R. Bailey, A. R. Sprenger, F. Grillo, H. Löwen, and L. Isa. Fitting an active brownian particle's mean-squared displacement with improved parameter estimation. *Physical Review E*, 106:L052602, Nov 2022. doi: 10.1103/PhysRevE.106.L052602.

V. Balakrishnan. The diffusion equation. In *Elements of Nonequilibrium Statistical Mechanics*, chapter 7, pages 81–96. Springer International Publishing, Basel, Switzerland, 2021. ISBN 978-3-030-62233-6. doi: 10.1007/978-3-030-62233-6_7. URL https://doi.org/10.1007/978-3-030-62233-6_7.

T. Ball, J. Protzenko, J. Bishop, M. Moskal, P. de Halleux, M. Braun, S. Hodges, and C. Riley. Microsoft touch develop and the bbc micro:bit. In *ICSE 2016 Companion*. ACM - Association for Computing Machinery, May 2016. URL https://www.microsoft.com/en-us/research/publication/microsoft-touch-develop-and-the-bbc-microbit/.

G. Ballestin, F. Solari, and M. Chessa. Perception and action in peripersonal space: A comparison between video and optical see-through augmented reality devices. *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 184–189, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00063.

P. Barr, R. Biddle, and J. Noble. A semiotic model of user-interface metaphor. In K. Liu, editor, *Virtual, Distributed and Flexible Organisations: Studies in Organisational Semiotics*, pages 189–215. Springer Netherlands, Dordrecht, Netherlands, 2004. doi: 10.1007/1-4020-2162-3_13.

L. Bartram, C. Ware, and T. Calvert. Moving Icons: Detection And Distraction. *Proceedings of the INTERACT '01: IFIP TC13 International Conference on Human-Computer Interaction*, pages 157–166, Jul 2001. URL `https://cir.nii.ac.jp/crid/1572261550800419328`.

P. Baudisch, N. Good, and P. Stewart. Focus plus context screens: Combining display technology with visualization techniques. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, UIST '01, page 31–40, New York, NY, 2001. Association for Computing Machinery. doi: 10.1145/502348.502354.

P. Baudisch, D. DeCarlo, A. T. Duchowski, and W. S. Geisler. Focusing on the Essential: Considering Attention in Display Design. *Communication of the ACM*, 46(3):60—-66, Mar 2003. doi: 10.1145/636772.636799.

V. Beiner, T. Gesslein, D. Schneider, F. Kawala, A. Otte, P. O. Kristensso, M. Michel Pahud, E. Ofek, C. Campos, M. Kljun, K. Č. Pucihar, and J. Grubert. PoVRPoint: Authoring Presentations in Mobile Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2069–2079, 2022. doi: 10.1109/TVCG.2022.3150474.

F. O. Black, C. Wall, H. E. Rockette, and R. Kitch. Normal Subject Postural Sway during the Romberg Test. *American Journal of Otolaryngology*, 3(5):309–318, 1982. doi: 10.1016/S0196-0709(82)80002-1.

Bootstrap. Introduction, 2023. URL `https://getbootstrap.com/docs/5.1/getting-started/introduction/`. Software.

R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. *Eurographics 2013 - State of the Art Reports*, 2013. ISSN 1017-4656. doi: 10.2312/conf/EG2013/stars/039-063.

L. Bottou. Stochastic Gradient Descent Tricks. In G. Montavon, G. B. Orr, & K.-R. Müller, editor, *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pages 421–436. Springer-Verlag Berlin Heidelberg, Berlin, German, 2012. ISBN 978-3-642-35289-8.

R. Brath. 3D InfoVis is here to stay: Deal with it. *2014 IEEE VIS International Workshop on 3DVis (3DVis)*, pages 25–31, 2014. doi: 10.1109/3DVis.2014.7160096.

B. F. Braumoeller. Hypothesis testing and multiplicative interaction terms. *International Organization*, 58(4):807–820, 2004. doi: 10.1017/S0020818304040251.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.

W. H. Brekhus, J. F. Galliher, and J. F. Gubrium. The Need for Thin Description. *Qualitative Inquiry*, 11(6):861–879, 2005. doi: 10.1177/1077800405280663.

N. Bressa, H. Korsgaard, A. Tabard, S. Houben, and J. Vermeulen. What's the Situation with Situated Visualization? A Survey and Perspectives on Situatedness. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):107–117, 2022. doi: 10.1109/TVCG.2021.3114835.

H. Brody, M. R. Rip, P. Vinten-Johansen, N. Paneth, and S. Rachman. Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet*, 356(9223):64–68, Jul 2000. doi: 10.1016/S0140-6736(00)02442-9.

J. Brooke. SUS: 'A quick and dirty' usability scale. In P. W. Jordan, B. Thomas, I. L. McClelland & B. Weerdmeester, editor, *Usability Evaluation In Industry*, volume 189. CRC Press, Boca Raton, FL, 11 1996.

V. Bruan and V. Clarke. Thematic analysis. In *APA handbook of research methods in psychology*, volume 2, pages 57–71. American Psychological Association, 2012. doi: doi.org/10.1037/13620-004.

S. Burigat, L. Chittaro, and S. Gabrielli. Visualizing locations of off-screen objects on mobile devices: a comparative evaluation of three approaches. *MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 239 – 246, Sep 2006. doi: 10.1145/1152215.1152266.

W. Büschel, S. Vogt, and R. Dachselt. Augmented reality graph visualizations. *IEEE Computer Graphics and Applications*, 39(3):29–40, 2019. doi: 10.1109/MCG.2019.2897927.

N. Cao, Y.-R. Lin, D. Gotz, and F. Du. Z-Glyph: Visualizing outliers in multivariate data. *Information Visualization*, 17(1):22–40, 2018. doi: 10.1177/1473871616686635.

D. Chandler. Models of the Sign. In *Semiotics: The Basics*, chapter 1, pages 11–67. Routledge, New York, NY, 3rd edition, 2018. ISBN 978-1-315-31105-0.

Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu, and Y. Wu. MARVisT: Authoring Glyph-Based Visualization in Mobile Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2645–2658, 2020. doi: 10.1109/TVCG.2019.2892415.

G. K. Cheung, T. Zimmermann, and N. Nagappan. The First Hour Experience: How the Initial Play Can Engage (or Lose) New Players. *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, pages 57—66, 2014. doi: 10.1145/2658537.2658540.

City of Toronto. Apartment Building Registration, 2021. URL `https://open.toronto.ca/dataset/apartment-building-registration/`. Data Set.

G. Clifford. Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures: Selected essays*, volume 3, pages 5–6. Basic Books, 1973. ISBN 9780465093564, 0465093566.

A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys*, 41(1), Jan 2009. doi: 10.1145/1456650.1456652.

E. A. Codling, M. J. Plank, and S. Benhamou. Random walk models in biology. *Journal of The Royal Society Interface*, 5(25):813–834, 2008. doi: 10.1098/rsif.2008.0014.

J. Cohen. The Analysis of Variance and Covariance. In *Statistical Analysis for the Behavioral Sciences*, chapter 8, pages 407–465. Routledge, New York, NY, 1988. doi: 10.4324/9780203771587.

D. S. Collingridge. A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, 7(1):81–97, 2013. doi: 10.1177/1558689812454457.

A. Çöltekin, A. L. Griffin, A. Slingsby, A. C. Robinson, S. Christophe, V. Rautenbach, M. Chen, C. Pettit, and A. Klippel. Geospatial Information Visualization and Extended Reality Displays. In H. Guo, M. F. Goodchild & A. Annoni, editor, *Manual of Digital Earth*, chapter 7, pages 229–277. Springer Singapore, Singapore, 2020. ISBN 978-981-32-9915-3. doi: 10.1007/978-981-32-9915-3_7.

A. Comber, C. Brunsdon, M. Charlton, G. Dong, R. Harris, B. Lu, Y. Lü, D. Murakami, T. Nakaya, Y. Wang, and P. Harris. A route map for successful applications of geographically weighted regression. *Geographical Analysis*, 55(1):155–178, 2023. doi: 10.1111/gean.12316.

K. Coninx, F. Van Reeth, and E. Flerackers. A hybrid 2d/3d user interface for immersive object modeling. *Proceedings Computer Graphics International*, pages 47–55, 1997. doi: 10.1109/CGI.1997.601270.

F. Crameri, G. E. Sheperd, and P. J. Heron. The misuse of colour in science communication. *Nature Communications*, 11(5444 (2020)), October 2020. doi: 10.1038/s41467-020-19160-7.

G. Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, 2014. doi: 10.1177/0956797613504966.

J. I. Daoud. Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, 949, Dec 2017. doi: 10.1088/1742-6596/949/1/012009.

J. Demšar, T. Curk, A. Erjavec, v. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14(1):2349–2353, Jan 2013. ISSN 1532-4435.

T. Dewhirst and W. B. Lee. Cigarette advertising in the republic of korea: a case illustration of the one. *Tobacco Control*, 21(6):584–588, 2012. doi: 10.1136/tobaccocontrol-2011-050315.

S. Di Verdi, D. Nurmi, and T. Hollerer. ARWin - a desktop augmented reality Window Manager. *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 298–299, 2003. doi: 10.1109/ISMAR.2003.1240729.

A. Din. Visualizing and Comparing Residential Permit Data Using Lollipop Plots. *Cityscape*, 21(2):175–178, 2019. URL `https://www.jstor.org/stable/26696382`.

A. Dix. Designing for Appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but Not as We Know It - Volume 2*, BCS-HCI '07, page 27–30, Swindon, GBR, 2007. BCS Learning & Development Ltd. ISBN 9781902505954.

E. Drewyer. NativeWebSocket, 2023. URL `https://github.com/endel/NativeWebSocket`. Software.

T. Drey, J. Gugenheimer, J. Karlbauer, M. Milo, and E. Rukzio. Vrsketchin: Exploring the design space of pen and tablet interaction for 3d sketching in virtual reality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–14, 2020. doi: 10.1145/3313831.3376628.

J. Du, Y. Shi, Z. Zou, and D. Zhao. Covr: Cloud-based multiuser virtual reality headset system for project communication of remote users. *Journal of Construction Engineering and Management*, 144(2):04017109, 2018. doi: 10.1061/(ASCE)CO.1943-7862.0001426.

J. J. Dudley and P. O. Kristensson. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2), Jun 2018. doi: 10.1145/3185517.

I. K. Duncan, S. Tingsheng, S. T. Perrault, and M. T. Gastner. Task-based effectiveness of interactive contiguous area cartograms. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2136–2152, 2021. doi: 10.1109/TVCG.2020.3041745.

M. A. Durivage. Log-normal distribution. In *Practical Engineering, Process, and Reliability Statistics (2nd Edition)*, chapter 27, pages 219–220. American Society for Quality (ASQ), Milwaukee, WI, 2022. ISBN 978-1-63694-015-1.

T. Dwyer, K. Marriott, T. Isenberg, K. Klein, N. Riche, F. Schreiber, W. Stuerzlinger, and B. H. Thomas. Immersive analytics: An introduction. In K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, & B. H. Thomas, editor, *Immersive Analytics*, pages 1–23. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-030-01388-2_1.

Elections Ontario. Graphics & charts, 2022. URL `https://results.elections.on.ca/en/graphics-charts`. Image.

L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests. page 754–768, 2021. doi: 10.1145/3472749.3474784.

B. Ens, B. Bach, M. Cordeil, U. Engelke, M. Serrano, W. Willett, A. Prouzeau, C. Anthes, W. Büschel, C. Dunne, T. Dwyer, J. Grubert, J. H. Haga, N. Kirshenbaum, D. Kobayashi, T. Lin, M. Olaosebikan, F. Pointecker, D. Saffo, N. Saquib, D. Schmalstieg, D. A. Szafir, M. Whitlock, and Y. Yang. Grand challenges in immersive analytics. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3446866.

A. Erickson, K. Kim, G. Bruder, and G. F. Welch. A Review of Visual Perception Research in Optical See-Through Augmented Reality. *ICAT-EGVE 2020 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 2020. doi: 10.2312/egve.20201256.

A. Erickson, K. Kim, A. Lambert, G. Bruder, M. P. Browne, and G. F. Welch. An extended analysis on the benefits of dark mode user interfaces in optical see-through head-mounted displays. *ACM Transactions on Applied Perception*, 18(3), May 2021. doi: 10.1145/3456874.

J. M. Evangelista Belo, A. M. Feit, T. Feuchtner, and K. Grønbæk. XRgonomics: Facilitating the Creation of Ergonomic 3D Interfaces. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445349.

S. Feiner and A. Shamash. Hybrid User Interfaces: Breeding Virtually Bigger Interfaces for Physically Smaller Computers. *Proceedings of the 4th Annual ACM Symposium on User Interface Software and Technology*, page 9–17, 1991. doi: 10.1145/120782.120783.

S. Feiner, B. MacIntyre, M. Haupt, and E. Solomon. Windows on the World: 2D Windows for 3D Augmented Reality. *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology*, page 145–155, 1993. doi: 10.1145/168642.168657.

V. Ferrer, Y. Yang, A. Perdomo, and J. Quarles. Consider your clutter: Perception of virtual object motion in AR. *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013. doi: 10.1109/ISMAR.2013.6671835.

J. P. Freiwald, N. Katzakis, and F. Steinicke. Camera time warp: Compensating latency in video see-through head-mounted-displays for reduced cybersickness effects. *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, (9), 2018. doi: 10.1145/3281505.3281521.

O. Friard and M. Gamba. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 7 (11):1325–1330, 2016. doi: https://doi.org/10.1111/2041-210X.12584.

R. J. Friedrich. In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, 26(4):797–833, 1982. URL http://www.jstor.org/stable/2110973.

H. Fritz, L. A. García-Escudero, and A. Mayo-Iscar. tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12), 2012. doi: 10.18637/jss.v047.i12.

J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim. A Systematic Review of Experimental Studies on Data Glyphs. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1863–1879, Jul 2017. doi: 10.1109/TVCG.2016.2549018.

H. G. Funkhouser and H. M. Walker. Playfair and his charts. *Economic History*, (10): 103–109, 1935. ISSN 27541096. URL http://www.jstor.org/stable/45366440.

A. Getis. Spatial autocorrelation. In M. M. Fischer & A. Getis, editor, *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pages 255–278. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi: 10.1007/978-3-642-03647-7_14.

C. D. Ghilani. Confidence interval. In *Adjustment Computations - Spatial Data Analysis (6th Edition)*, chapter 4, pages 57–78. John Wiley & Sons, 2018. ISBN 978-1-119-38598-1. URL https://app.knovel.com/hotlink/khtml/id:kt011UB6TR/adjustment-computations/confidence-interval-population.

G. Glatting, P. Kletting, S. N. Reske, K. Hohl, and C. Ring. Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test. *Medical Physics*, 34(11):4285–4292, 2007. doi: 10.1118/1.2794176.

Google. Google Earth (Desktop Version), 2023. URL https://www.google.com/earth/about/versions/. Software.

I. Gołębiowska and A. Çöltekin. What's wrong with the rainbow? an interdisciplinary review of empirical evidence for and against the rainbow color scheme in visualizations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:195–208, 2022. doi: 10.1016/j.isprsjprs.2022.10.002.

D. S. Grebenkov. Probability distribution of the time-averaged mean-square displacement of a gaussian process. *Physics Review E*, 84:031124, Sep 2011. doi: 10.1103/PhysRevE.84.031124.

R. Gruen, E. Ofek, A. Steed, R. Gal, M. Sinclair, and M. Gonzalez-Franco. Measuring System Visual Latency through Cognitive Latency on Video See-Through AR devices. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 791–799, 2020. doi: 10.1109/VR46266.2020.00103.

U. Gruenefeld, D. Ennenga, A. E. Ali, W. Heuten, and S. Boll. EyeSee360: designing a visualization technique for out-of-view objects in head-mounted augmented reality. *Proceedings of the 5th Symposium on Spatial User Interaction - SUI '17*, pages 109–118, Oct 2017. doi: 10.1145/3131277.3132175.

U. Gruenefeld, D. Lange, L. Hammer, S. Boll, and W. Heuten. FlyingARrow: Pointing Towards Out-of-View Objects on Augmented Reality Devices. *Proceedings of the 7th ACM International Symposium on Pervasive Displays - PerDis '18*, pages 1–6, Apr 2018. doi: 10.1145/3205873.3205881.

C. Grégoire. Pattern-of-Life Analysis. In *A Theory of the Drone.*, chapter 5. The New Press, New York, NY, 2013. ISBN 9781595589750.

A. K. Gupta. Beta distribution. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 144–145. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-04898-2_144.

S. M. Göbel. Up or down? Reading direction influences vertical counting direction in the horizontal plane – a cross-cultural comparison. *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.00228.

M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042.

S. G. Hart. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909.

Health Canada. Guidelines for Canadian Drinking Water Quality: Guideline Technical Document – Colour, 1995. URL `https://www.canada.ca/en/health-canada/services/publications/healthy-living/guidelines-canadian-drinking-water-quality-guideline-technical-document-colour.html`.

R. Herriott. What kind of research is research through design. Manchester, United Kingdoms, Sep 2019. URL `https://adk.elsevierpure.com/en/publications/what-kind-of-research-is-research-through-design`.

D. C. Hoaglin. John W. Tukey and Data Analysis. *Statistical Science*, 18(3):311–318, 2003. URL `http://www.jstor.org/stable/3182748`.

J. L. Hodges Jr. and E. L. Lehmann. Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics*, 34(2):598 – 611, 1963. doi: 10.1214/aoms/1177704172.

B. D. Hoffman, G. Massiera, K. M. V. Citters, and J. C. Crocker. The consensus mechanics of cultured mammalian cells. *Proceedings of the National Academy of Sciences*, 103(27):10259–10264, 2006. doi: 10.1073/pnas.0510348103.

T. Holz, A. G. Campbell, G. M. O'Hare, J. W. Stafford, A. Martin, and M. Dragone. Mira—mixed reality agents. *International Journal of Human-Computer Studies*, 69(4):251–268, 2011. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2010.10.001.

D. Hosken, D. Buss, and D. Hodgson. Beware the f test (or, how to compare variances). *Animal Behaviour*, 136:119–126, 2018. doi: 10.1016/j.anbehav.2017.12.014.

S. Hu, J. Malloch, and D. Reilly. A Comparative Evaluation of Techniques for Locating Out-of-View Targets in Virtual Reality. pages 202 – 212, 2021. doi: 10.20380/GI2021.32.

S. H. Hu, D. Reilly, and S. Bashbaghi. Augmented Reality + Tablet Interface for Model Selection, 2022. Patent No. PCT/IB2022/052779 [Pending].

S. Hubenschmid, J. Zagermann, S. Butscher, and H. Reiterer. STREAM: Exploring the Combination of Spatially-Aware Tablets with Augmented Reality Head-Mounted Displays for Immersive Analytics. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445298.

C. Hull and W. Willett. Building with Data: Architectural Models as Inspiration for Data Physicalization. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 1217–1264, 2017. doi: 10.1145/3025453.3025850.

E. L. Hutchins, J. D. Hollan, and D. A. Norman. Direct Manipulation Interfaces. *Human–Computer Interaction*, 1(4):311–338, 1985. doi: 10.1207/s15327051hci0104_2.

Immigration, Refugee, and Citizenship Canada. Prince Edward Island (PE) - Facts, Flags and Symbols. In *Celebrate being Canadian*. The Government of Canada, August 2020. URL `https://www.canada.ca/en/immigration-refugees-citizenship/services/canadians/celebrate-being-canadian.html`. Image.

A. Isoni. Supervised Machine Learning. In *Machine Learning for the Web*, chapter 3, pages 73–118. Packt Publishing, Birmingham, UK, 1st edition, 7 2016.

Y. Itoh and G. Klinker. Light-field correction for spatial calibration of optical see-through head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 21(4):471–480, 2015. doi: 10.1109/TVCG.2015.2391859.

Y. Itoh, T. Langlotz, J. Sutton, and A. Plopski. Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays. *ACM Comput. Surv.*, 54(6), Jul 2021. doi: 10.1145/3453157.

T. Jankun-Kelly, Y. Lanka, and J. Swan II. An Evaluation of Glyph Perception for Real Symmetric Traceless Tensor Properties. *Computer Graphics Forum*, 29(3): 1133–1142, 2010. doi: 10.1111/j.1467-8659.2009.01711.x.

H. Jenkins-Smith, J. Ripberger, G. Copeland, M. Nowlin, T. Hughes, A. Fister, and W. Wehde. Quantitative Research Methods for Political Science, Public Policy and Public Administration: 4th Edition With Applications in R. chapter 13, pages 127–133. University of Oklahoma, Norman, OK, 2021. URL `https://bookdown.org/josiesmith/qrmbook/multiple-regression-and-model-building.html#model-building`.

J. E. Johnston, K. J. Berry, and J. Paul W. Mielke. Measures of Effect Size for Chi-Squared and Likelihood-Ratio Goodness-of-Fit Tests. *Perceptual and Motor Skills*, 103(2):412–414, 2006. doi: 10.2466/pms.103.2.412-414.

D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks. Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020. doi: 10.1016/j.cirpj.2020.02.002.

K. Lebedev, et al. SortableJS, 2022. URL `https://github.com/SortableJS/Sortable`. Software.

Kcida10. File:HoloLens 2.jpeg, 2015. URL `https://commons.wikimedia.org/wiki/File:HoloLens_2.jpeg`. Image.

D. S. Kerby. The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. *Comprehensive Psychology*, 3:11.IT.3.1, 2014. doi: 10.2466/11.IT.3.1.

B. Keskin. Multilevel approach to the analysis of housing submarkets. *Regional Studies, Regional Science*, 9(1):264–279, 2022. doi: 10.1080/21681376.2022.2067005.

D. Kharel. Visual Ethnography, Thick Description and Cultural Representation. *Dhaulagiri Journal of Sociology and Anthropology*, 9:147–160, Dec 2015. doi: 10.3126/dsaj.v9i0.14026.

K. Kim, A. Erickson, A. Lambert, G. Bruder, and G. Welch. Effects of Dark Mode on Visual Fatigue and Acuity in Optical See-Through Head-Mounted Displays. *Symposium on Spatial User Interaction*, 2019. doi: 10.1145/3357251.3357584.

P. Kovesi. Bad Colour Maps Hide Big Features and Create False Anomalies. *ASEG Extended Abstracts*, 2015(1):1–4, 2015. doi: 10.1071/ASEG2015ab107.

P. Kovesi. Colour amplifies relief shading. *ASEG Extended Abstracts*, 2019(1):1–3, 2019. doi: 10.1080/22020586.2019.12073048.

B. C. Kress and W. J. Cummings. 11-1: Invited Paper: Towards the Ultimate Mixed Reality Experience: HoloLens Display Architecture Choices. *SID Symposium Digest of Technical Papers*, 48(1):127–131, 2017. doi: 10.1002/sdtp.11586.

E. Kruijff, J. E. Swan, and S. Feiner. Perceptual issues in augmented reality revisited. *2010 IEEE International Symposium on Mixed and Augmented Reality*, pages 3–12, 2010. doi: 10.1109/ISMAR.2010.5643530.

P. Kumar, L. K. Sharma, P. C. Pandey, S. Sinha, and M. S. Nathawat. Geospatial Strategy for Tropical Forest-Wildlife Reserve Biomass Estimation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2):917–923, 2013. doi: 10.1109/JSTARS.2012.2221123.

R. Langner, M. Satkowski, W. Büschel, and R. Dachselt. MARVIS: Combining Mobile Devices and Augmented Reality for Visual Data Analysis. *Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems*, 5 2021. doi: 10.1145/3411764.3445593.

C. W. Lau, Q. V. Nguyen, Z. Qu, S. Simoff, and D. Catchpoole. Immersive Intelligence Genomic Data Visualisation. *Proceedings of the Australasian Computer Science Week Multiconference*, 2019. doi: 10.1145/3290688.3290722.

D. Ledo, S. Houben, J. Vermeulen, N. Marquardt, L. Oehlberg, and S. Greenberg. Evaluation Strategies for HCI Toolkit Research. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018. doi: 10.1145/3173574.3173610.

S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics*, 23 (1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920.

J. R. Lewis and J. Sauro. Can I Leave This One Out? The Effect of Dropping an Item From the SUS. *Journal of Usability Studies*, 13(1), 2017. URL `https://dl.acm.org/doi/10.5555/3173069.3173073`.

J. R. Lewis, B. S. Utesch, and D. E. Maher. UMUX-LITE: When There's No Time for the SUS. *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 2099–2102, Apr 2013. doi: 10.1145/2470654.2481287.

M. S. Lewis-Beck and A. Skalaban. The R-Squared: Some Straight Talk. *Political Analysis*, 2:153–171, 1990. doi: 10.1093/pan/2.1.153.

M. J. Liberatore and W. P. Wagner. Virtual, mixed, and augmented reality: a systematic review for immersive systems research. *Virtual Reality*, 25(3):773–799, Sep 2021. doi: 10.1007/s10055-020-00492-0.

M. A. Livingston, J. H. Barrow, and C. M. Sibley. Quantification of contrast sensitivity and color perception using head-worn augmented reality displays. *2009 IEEE Virtual Reality Conference*, pages 115–122, 2009. doi: 10.1109/VR.2009.4811009.

A. D. Logvinenko. Does luminance contrast determine lightness? *Spatial Vision*, 18 (3):337–345, 2005. doi: 10.1163/1568568054089357.

H. Luepsen. The aligned rank transform and discrete variables: A warning. *Communications in Statistics - Simulation and Computation*, 46(9):6923–6936, 2017. doi: 10.1080/03610918.2016.1217014.

A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics and uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012. doi: 10.1109/TVCG.2012.279.

I. S. MacKenzie. Hypothesis Testing. In *Human-computer Interaction*, chapter 6, pages 191–232. Morgan Kaufmann, Boston, USA, 2013. ISBN 978-0-12-405865-1. doi: 10.1016/B978-0-12-405865-1.00006-6.

L. Magee. R2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253, 1990. URL `http://www.jstor.org/stable/2685352`.

Magic Leap, Inc. Four Optics Breakthroughs to Power Enterprise AR, 2022. URL `https://www.magicleap.com/hubfs/Magic-Leap-2-Optics-Highlights-Paper.pdf`. Announcement.

S. Maher and D. Spicer. Digital Earth Workbench: The Earth's Magnetic Field, November 1999. URL `https://svs.gsfc.nasa.gov/803`. Video.

Mapbox. Maps and location for developers, 2022. URL `https://www.mapbox.com/`. Software.

M. Marozzi. Testing for concordance between several criteria. *Journal of Statistical Computation and Simulation*, 84(9):1843–1850, 2014. doi: 10.1080/00949655.2013.766189.

G. Martynenko. Semiotics of Statistics. *Journal of Quantitative Linguistics*, 10(2): 105–115, 2003. doi: 10.1076/jqul.10.2.105.16712.

N. S. Mathews, S. Chimalakonda, and S. Jain. Air: An augmented reality application for visualizing air pollution. *2021 IEEE Visualization Conference (VIS)*, pages 146–150, 2021. doi: 10.1109/VIS49827.2021.9623287.

A. Mayorga and M. Gleicher. Splatterplots: overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, 19(9):1526–1538, Sep 2013. doi: 10.1109/TVCG.2013.65.

J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. A. S. Buxton & S. Greenberg, editor, *Readings in Human–Computer Interaction*, Interactive Technologies, pages 152–169. Morgan Kaufmann, Burlington, MA, 1995. doi: 10.1016/B978-0-08-051574-8.50019-4.

D. J. McLean and M. A. Skowron Volponi. trajr: An r package for characterisation of animal trajectories. *Ethology*, 124(6):440–448, 2018. doi: 10.1111/eth.12739.

L. McNabb and R. S. Laramee. Multivariate maps—a glyph-placement algorithm to support multivariate geospatial visualization. *Information*, 10(10), 2019. doi: 10.3390/info10100302.

A. Melamud, S. Hagstrom, and E. Traboulsi. Color vision testing. *Ophthalmic Genetics*, 25(3):159–187, Sept. 2004. doi: 10.1080/13816810490498341.

MetaMarket. File:Meta 2.jpg, 2016. URL https://commons.wikimedia.org/wiki/File:Meta_2.jpg. Image.

Microsoft. MRTK2, 2022. URL https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/?view=mrtkunity-2022-05. Software.

Microsoft. Surface book 3 - tech specs, 2022. URL https://www.microsoft.com/en-ca/surface/devices/surface-book-3/tech-specs.

H. J. Miller. Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94(2):284–289, 2004. doi: 10.1111/j.1467-8306.2004.09402005.x.

Miro. Miro, 2023. URL https://miro.com/. Software.

P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23, 1950. URL http://www.jstor.org/stable/2332142.

K. Moreland. Diverging Color Maps for Scientific Visualization. *Advances in Visual Computing*, pages 92–103, 2009. doi: 10.1007/978-3-642-10520-3_9.

Mozilla. FireFox, 2023. URL https://www.mozilla.org/en-CA/firefox/. Software.

H. Müller, R. Reihs, K. Zatloukal, and A. Holzinger. Analysis of biomedical data with multilevel glyphs. *BMC Bioinformatics*, 15(6):S5, May 2014. doi: 10.1186/1471-2105-15-S6-S5.

N. J. D. Nagelkerke. A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692, 1991. URL http://www.jstor.org/stable/2337038.

S. Nakagawa, P. C. D. Johnson, and H. Schielzeth. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, 14(134), 2017. doi: 10.1098/rsif.2017.0213.

E. Nas, F. Longhi, L. Terceiro, T. Azevedo, and T. Valente. Future visions for a decolonized future of hci. In C. Stephanidis, M/ Antona, S. Ntoa & G. Salvendy, editor, *HCI International 2023 Posters*, pages 109–116, Basel, Switzerland, 2023. Springer Nature Switzerland. doi: 10.17613/30bc-6j76.

Niantic. Pokemon Go, 2022. URL `https://pokemongolive.com/`. Software.

M. B. Nuijten, C. H. J. Hartgerink, M. A. L. M. van Assen, S. Epskamp, and J. M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, Dec 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0664-2.

J. R. Nuñez, C. R. Anderton, and R. S. Renslow. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLOS ONE*, 13(7):1–14, 08 2018. doi: 10.1371/journal.pone.0199239.

P. Ondrejka. Mapping election results in proportional electoral systems. *Journal of Maps*, 12(sup1):591–596, 2016. doi: 10.1080/17445647.2016.1239558.

T. Opach and J. K. Rød. Augmenting the usability of parallel coordinate plot: The polyline glyphs. *Information Visualization*, 17(2):108–127, 2018. doi: 10.1177/1473871617693041.

T. Opach, S. Popelka, J. Dolezalova, and J. K. Rød. Star and polyline glyphs in a grid plot and on a map display: which perform better? *Cartography and Geographic Information Science*, 45(5):400–419, 2018. doi: 10.1080/15230406.2017.1364169.

OpenJS Foundation and JS Query Contributors. jQuery, 2023. URL `https://jquery.com/`. Software.

Y. Othman, M. Khalaf, A. Ragab, A. Salaheldin, R. Ayman, and N. Sharaf. Eye-to-eye: Towards visualizing eye gaze data. *2020 24th International Conference Information Visualisation (IV)*, pages 729–733, 2020. doi: 10.1109/IV51561.2020.00128.

F. Paas, A. Renkl, and J. Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4, 2003. doi: 10.1207/S15326985EP3801_1.

A. Palén. JavaScript functions to calculate combinations of elements in Array, 2012. URL `https://gist.github.com/axelpale/3118596`. Software.

Y. Park, K. Sharkey, and K. Eveleigh. Designing content for holographic display. Microsoft, May 2021. URL https://learn.microsoft.com/en-us/windows/mixed-reality/design/designing-content-for-holographic-display.

D. S. Parker, E. Congdon, and R. M. Bilder. Hypothesis exploration with visualization of variance. *BioData Mining*, 7, Jul 2014. doi: 10.1186/1756-0381-7-11. URL https://pubmed.ncbi.nlm.nih.gov/25097666.

L. Pavanatto, C. North, D. A. Bowman, C. Badea, and R. Stoakley. Do we still need physical monitors? An evaluation of the usability of AR virtual monitors for productivity work. *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 759–767, 2021. doi: 10.1109/VR50410.2021.00103.

Z. L. Peichao Gao and Z. Qin. Usability of value-by-alpha maps compared to area cartograms and proportional symbol maps. *Journal of Spatial Science*, 64(2):239–255, 2019. doi: 10.1080/14498596.2018.1440649.

C. Perin, P. Dragicevic, and J. D. Fekete. Revisiting Bertin Matrices: New Interactions for Crafting Tabular Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2082–2091, Dec 2014. doi: 10.1109/TVCG.2014.2346279.

E. Persson. Indoors skyboxes, 2013. URL https://opengameart.org/content/indoors-skyboxes. Game Assets.

R. A. Peterson and J. E. Cavanaugh. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 47(13-15):2312–2327, 2020. doi: 10.1080/02664763.2019.1630372.

S. D. Peterson, M. Axholt, M. Cooper, and S. R. Ellis. Visual clutter management in augmented reality: Effects of three label separation methods on spatial judgments. *2009 IEEE Symposium on 3D User Interfaces*, pages 111–118, 2009. doi: 10.1109/3DUI.2009.4811215.

V. Peña-Araya, E. Pietriga, and A. Bezerianos. A Comparison of Visualizations for Identifying Correlation over Space and Time. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):375–385, 2020. doi: 10.1109/TVCG.2019.2934807.

T. M. Porter. Thin description: Surface and depth in science and science studies. *Osiris*, 27(1):209–226, 2012. doi: 10.1086/667828.

M. Poupard, M. Ferrari, J. Schluter, R. Marxer, P. Giraudet, V. Barchasz, V. Gies, G. Pavan, and H. Glotin. Real-time passive acoustic 3d tracking of deep diving cetacean by small non-uniform mobile surface antenna. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8251–8255, 2019. doi: 10.1109/ICASSP.2019.8683883.

K. Pulli. 11-2: Invited paper: Meta 2: Immersive optical-see-through augmented reality. *SID Symposium Digest of Technical Papers*, 48(1):132–133, 2017. doi: 10.1002/sdtp.11588.

Pulpil Labs. hmd-eyes, 2021. URL `https://github.com/pupil-labs/hmd-eyes`. Game Assets.

P. S. Quinan, L. Padilla, S. H. Creem-Regehr, and M. Meyer. Hue Bands and Human Perception: Revisiting the Rainbow. *Proceedings of the IEEE Information Visualization Conference - Posters (InfoVis)*, 2017. URL `https://vdl.sci.utah.edu/publications/2017_infovis_huebands/`.

I. M. Ratner and J. Harvey. Vertical Slicing: Smaller is Better. *2011 Agile Conference*, pages 240–245, 2011. doi: 10.1109/AGILE.2011.46.

K. Reda and D. A. Szafir. Rainbows revisited: Modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1032–1042, 2021. doi: 10.1109/TVCG.2020.3030439.

D. Reilly and B. MacKay. Annotating Ecology: Looking to Biological Fieldwork for Mobile Spatial Annotation Workflows. *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, page 35–44, 2013. doi: 10.1145/2493190.2493245.

J. Rekimoto and Y. Ayatsuka. CyberCode: Designing Augmented Reality Environments with Visual Tags. *DARE '00: Proceedings of DARE 2000 on Designing augmented reality environments*, 2000. doi: 10.1145/354666.354667.

J. D. Rights and S. K. Sterba. New Recommendations on the Use of R-Squared Differences in Multilevel Model Comparisons. *Multivariate Behavioral Research*, 55(4):568–599, 2020. doi: 10.1080/00273171.2019.1660605. PMID: 31559890.

A. Rocha, U. Alim, J. D. Silva, and M. C. Sousa. Decal-maps: Real-time layering of decals on surfaces for multivariate visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):821–830, 2017. doi: 10.1109/TVCG.2016.2598866.

T. Ropinski, S. Oeltze, and B. Preim. Survey of glyph-based visualization techniques for spatial multivariate medical data. *Computers & Graphics*, 35(2):392–401, 2011. doi: 10.1016/j.cag.2011.01.011.

R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision*, 7 (2):17–17, Aug 2007. doi: 10.1167/7.2.17.

G. K. Rosenkranz. A note on the Hodges-Lehmann estimator. *Pharmaceutical Statistics*, 9(2):162–167, Apr 2010. doi: 10.1002/pst.387.

J. A. Rosenthal. Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research*, 21(4):37–59, 1996. doi: 10.1300/J079v21n04_02.

R. E. Roth, A. W. Woodruff, and Z. F. Johnson. Value-by-alpha maps: An alternative technique to the cartogram. *The Cartographic Journal*, 47(2):130–140, May 2010. doi: 10.1179/000870409X12488753453372.

D. Rouan. Parallax. In M. Gargaud, W. M. Irvine, R. Amils, H. J. Cleaves, D. L. Pinti, J. C. Quintanilla, D. Rouan, T. Spohn, S. Tirard, & M. Viso, editor, *Encyclopedia of Astrobiology*, pages 1837–1838. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. doi: 10.1007/978-3-662-44185-5_1153.

D. J. Salkeld and M. F. Antolin. Ecological fallacy and aggregated data: A case study of fried chicken restaurants, obesity and lyme disease. *EcoHealth*, 17(1):4–12, Mar 2020. doi: 10.1007/s10393-020-01472-1.

D. Salomon. Color. In *The Computer Graphics Manual*, pages 975–1003. Springer London, London, United Kingdoms, 2011. doi: 10.1007/978-0-85729-886-7_21.

A. C. Sampanes, P. Tseng, and B. Bridgeman. The role of gist in scene recognition. *Vision Research*, 48(21):2275–2283, 2008. doi: 10.1016/j.visres.2008.07.011.

K. A. Satriadi, J. Smiley, B. Ens, M. Cordeil, T. Czauderna, B. Lee, Y. Yang, T. Dwyer, and B. Jenny. Tangible Globes for Data Visualisation in Augmented Reality. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022. doi: 10.1145/3491102.3517715.

D. Schreuder. The mathematics of luminance. In *Outdoor Lighting: Physics, Vision and Perception*, pages 109–140. Springer Netherlands, Dordrecht, Nehterlands, 2008. ISBN 978-1-4020-8602-1. doi: 10.1007/978-1-4020-8602-1_4.

T. Sebeok. Semiotics and Linguistics. In S. P. X. Battestini, editor, *Developments in Linguistics and Semiotics Language Teaching and Learning Communication across Cultures*, Georgetown University Round Table on Languages and Linguistics, pages 1–18. Georgetown University, Washington, DC, 1986. ISBN 0-87840-121-0.

S. Shaki and M. H. Fischer. Deconstructing spatial-numerical associations. *Cognition*, 175:109–113, 2018. doi: 10.1016/j.cognition.2018.02.022.

O. Sheffet. Differentially Private Ordinary Least Squares. *Proceedings of the 34th Conference on Machine Learning*, 2017. URL `http://proceedings.mlr.press/v70/sheffet17a/sheffet17a.pdf`.

B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307.

M. J. Simpson. Mini-review: Far peripheral vision. *Vision Research*, 140:96–105, 2017. ISSN 0042-6989. doi: 10.1016/j.visres.2017.08.001.

Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila. A Survey on Mobile Augmented Reality With 5G Mobile Edge Computing: Architectures, Applications, and Technical Aspects. *IEEE Communications Surveys Tutorials*, 23(2):1160–1192, 2021. doi: 10.1109/COMST.2021.3061981.

N. Siva, A. Chaparro, and E. Palmer. Human Factors Principles Underlying Glyph Design: A Review of the Literature and an Agenda for Future Research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1659–1663, 2012. doi: 10.1177/1071181312561332.

A. Skupin and S. I. Fabrikant. Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2):99–119, 2003. doi: 10.1559/152304003100011081.

M. Snipes and D. C. Taylor. Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3(1):3–9, 2014. doi: 10.1016/j.wep.2014.03.001.

I. Soares, R. B. Sousa, M. Petry, and A. P. Moreira. Accuracy and Repeatability Tests on HoloLens 2 and HTC Vive. *Multimodal Technologies and Interaction*, 5 (8), 2021. ISSN 2414-4088. doi: 10.3390/mti5080047.

I. Spence. No Humble Pie: The Origins and Usage of a Statistical Chart. *Journal of Educational and Behavioral Statistics*, 30(4):353–368, 2005. doi: 10.3102/10769986030004353.

St. John's International Airport. St. John's International Airport Departures Lounge 2nd Floor, 2019. URL `https://stjohnsairport.com/wp-content/uploads/2019/09/YYT-Map-interior-SecondFloor-Sept2019-e1568812696234.png`. Image.

A. J. Stapleton. Research as Design-Design as Research. *DiGRA 2005 - Proceedings of the 2005 DiGRA International Conference: Changing Views: Worlds in Play*, 3, 2005. URL `http://www.digra.org/wp-content/uploads/digital-library/06278.40383.pdf`.

M. Stoelzle and L. Stein. Rainbow color map distorts and misleads research in hydrology – guidance for better visualizations and science communication. *Hydrology and Earth System Sciences*, 25(8):4549–4565, 2021. doi: 10.5194/hess-25-4549-2021.

M. Studer, G. Ritschard, A. Gabadinho, and N. S. Müller. Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3):471–510, 2011. doi: 10.1177/0049124111415372.

T.-L. Sun and W.-L. Kuo. Visual exploration of production data using small multiples design with non-uniform color mapping. *Computers & Industrial Engineering*, 43 (4):751–764, 2002. doi: 10.1016/S0360-8352(02)00137-7.

H. B. Surale, A. Gupta, M. Hancock, and D. Vogel. Tabletinvr: Exploring the design space for using a multi-touch tablet in virtual reality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–13, 2019. doi: 10.1145/3290605.3300243.

D. A. Szafir. The Good, the Bad, and the Biased: Five Ways Visualizations Can Mislead (and How to Fix Them). *Interactions*, 25(4):26—-33, Jun 2018. doi: 10.1145/3231772.

A. Søgaard. Compound constructions: A reply to Bundgaard et al. *Semiotica: Journal of the International Association for Semiotic Studies*, 2008(169):189–195, 2008. doi: 10.1515/SEM.2008.031.

Tableau. Tableau (version. 9.1). *Journal of the Medical Library Association : JMLA*, 104(2):182–183, Apr 2016. doi: 10.3163/1536-5050.104.2.022. Software.

The Government of Nova Scotia. Windosr / West Hants Together, 2020. URL `https://www.strongerregion.ca/index.php`. Website.

The Government of Nova Scotia. Nova Scotia Lake Chemistry Data, 2021. URL `https://data.novascotia.ca/Environment-and-Energy/Nova-Scotia-Lake-Chemistry-Data/vn55-yjyi`. Data Set.

A. Trigo. pagePiling.js, 2023. URL `https://alvarotrigo.com/pagePiling/`. Software.

M. Tönnis, D. A. Plecher, and G. Klinker. Representing information – Classifying the Augmented Reality presentation space. *Computers & Graphics*, 37(8):997–1011, 2013. doi: 10.1016/j.cag.2013.09.002.

Unity. ShaderLab command: ColorMask. In *Unity User Manual 2023.1*. Unity, January 2023. URL `https://docs.unity3d.com/2023.1/Documentation/Manual/SL-ColorMask.html`.

Unity. Unity, 2023. URL `https://unity.com/`. Software.

V. Agafonkin and Leaflet maintainers. Leaflet.js, 2023. URL `https://leafletjs.com/`. Software.

S. van den Elzen and J. J. van Wijk. Small Multiples, Large Singles: A New Approach for Visual Data Exploration. *Computer Graphics Forum*, 32(3pt2):191–200, 2013. doi: 10.1111/cgf.12106.

C. L. S. Veldkamp, M. B. Nuijten, L. Dominguez-Alvarez, M. A. L. M. van Assen, and J. M. Wicherts. Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLOS ONE*, 9(12):1–19, 12 2014. doi: 10.1371/journal.pone.0114876.

W3C. Compositing and Blending Level 1. In R. Cabanier & N. Andronikos, editor, *W3C standards and drafts*. W3C, 2015. URL `https://www.w3.org/TR/compositing-1/`. 2021-09-28.

G. Walsh, N. Andersen, N. Stoianov, and S. Jänicke. A survey of geospatial-temporal visualizations for military operations. *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - IVAPP*, pages 115–129, 2023. doi: 10.5220/0011902500003417.

M. O. Ward. Multivariate Data Glyphs: Principles and Practice. In *Handbook of Data Visualization*, pages 179–198. Springer Berlin Heidelberg, Berlin, Germany, 2008. doi: 10.1007/978-3-540-33037-0_8.

K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, page 7–9, 2019. doi: 10.1145/3308532.3329441.

K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 15(2):87–98, Jun 2021. doi: 10.1007/s12193-020-00332-0.

Q. Wen, C. Cheles, A. Buck, T. P. Milligan, V. Tieto, and C. McClister. QR code tracking overview. *Mixed Reality documentation*, Oct 2021. URL `https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/qr-code-tracking-overview`. Software.

M. Whitlock, K. Wu, and D. A. Szafir. Designing for mobile and immersive visual analytics in the field. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):503–513, 2020. doi: 10.1109/TVCG.2019.2934282.

W. Willett, B. Jenny, T. Isenberg, and P. Dragicevic. Lightweight Relief Shearing for Enhanced Terrain Perception on Interactive Maps. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3563—-3572, 2015. doi: 10.1145/2702123.2702172.

A. S. Williams and F. R. Ortega. Using a 6 degrees of freedom virtual reality input device with an augmented reality headset in a collaborative environment. *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 205–209, 2021. doi: 10.1109/VRW52623.2021.00045.

J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, pages 143–146, May 2011. doi: 10.1145/1978942.1978963.

H. Yamaguchi, M. Kitani, and H. Murakami. Robust testing procedures for scale differences in paired data. *Journal of Statistical Computation and Simulation*, 93 (12):1899–1923, 2023. doi: 10.1080/00949655.2022.2163645.

Y. Yang, T. Dwyer, K. Marriott, B. Jenny, and S. Goodwin. Tilt Map: Interactive Transitions Between Choropleth Map, Prism Map and Bar Chart in Immersive Environments. *IEEE Transactions on Visualization and Computer Graphics*, 27 (12):4507–4519, 2021. doi: 10.1109/TVCG.2020.3004137.

Y. Yang, W. Xia, F. Lekschas, C. Nobre, R. Krüger, and H. Pfister. The pattern is in the details: An evaluation of interaction techniques for locating, searching, and contextualizing details in multivariate matrix visualizations. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, (84), 2022. doi: 10.1145/3491102.3517673.

J. Zasdani. Hand-made raised-relief map of the High Tatras in scale 1: 50 000, 2007. URL `https://en.wikipedia.org/wiki/Terrain_cartography#/media/File:Tatry_Mapa_Plastyczna.JPG`. Image.

H. Zemanek. Semiotics and Programming Languages. *Communications of the ACM*, 9(3):139–143, Mar 1966. doi: 10.1145/365230.365249.

H. Zhang and N. K. Tripathi. Geospatial hot spot analysis of lung cancer patients correlated to fine particulate matter (PM2.5) and industrial wind in Eastern Thailand. *Journal of Cleaner Production*, 170:407–424, 2018. doi: 10.1016/j.jclepro.2017.09.185.

L. Zhang, Z. Ma, and L. Guo. An Evaluation of Spatial Autocorrelation and Heterogeneity in the Residuals of Six Regression Models. *Forest Science*, 55(6):533–548, 12 2009. doi: 10.1093/forestscience/55.6.533.

Z. Zhao, E. Zgraggen, L. De Stefani, C. Binnig, E. Upfal, and T. Kraska. Safe visual data exploration. *Proceedings of the 2017 ACM International Conference on Management of Data*, page 1671–1674, 2017. doi: 10.1145/3035918.3058749.

L. R. Zientek and Z. E. Yetkiner. Pooled Variance. In N. J. Salkind, editor, *Encyclopedia of Research Design*. SAGE Publications, 2010. doi: 10.4135/9781412961288.

S. T. Ziliak. Retrospectives: Guinnessometrics: The Economic Foundation of "Student's" t. *Journal of Economic Perspectives*, 22(4):199–216, Dec 2008. doi: 10.1257/jep.22.4.199.

S. T. Ziliak. How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough. *The American Statistician*, 73(sup1):281–290, 2019. doi: 10.1080/00031305.2018.1514325.

J. Zlatev. Cognitive semiotics. In P. P. Trifonas, editor, *International Handbook of Semiotics*, pages 1043–1067. Springer Netherlands, Dordrecht, Netherlands, 2015. ISBN 978-94-017-9404-6. doi: 10.1007/978-94-017-9404-6_47.

Appendix A

Multifactor Testing Procedures for Trajectory Data with Random Walk

This chapter is not a part of the main research. Rather, it serves two main purposes. The first one is to provide a chance to apply the knowledge of variance gained during the exploratory research. Since we decided to adopt likelihood as a measure of goodness-of-fit, much research on variance no longer fits in the main text. Secondly, it provides a preliminary treatise on specific types of trajectory analyses, and a further elaboration of the methods used in the synoptic study (Ch. 6). The work here could further be expanded, but it would be beyond the scope of this thesis. By including this chapter in a publicly available document (e.g. a thesis), we hope that it may still be useful for any researcher who performs similar analytical work.

After having performed multiple trajectory analyses in this thesis, we note there is a lack of literature on trajectory analysis with random walk. Therefore, we develop this appendix chapter as a summary of our trajectory analysis methods. Furthermore, we discuss the statistical underpinning of our work, and additional guidelines for future research. The guidelines provided here are not absolute; after all, there are myriad types of trajectories and researchers should not treat them all the same.

Trajectory data can arise from many types of human-computer interaction (HCI) studies. For example, in a Fitts's Law study, the participants moved a stylus from one point to another–creating trajectories between two points. In a more modern example in mixed reality, we may ask the participants to navigate around a large area to test a virtual environment. As long as there are a series of movements, there will be trajectories. Testing if study conditions affect trajectory data can be a challenging enterprise if the data contain random walk. Random walk arises when the participants have the freedom to move. In a Fitts's Law study, there is not much random walk since the starting point and the destination points are well-defined. However, if the participants can move freely, their paths can become unpredictable.

We developed a new testing procedure for trajectory data with random walk: posNOVA. While there are many techniques for dealing with trajectory data, they do not cope well with random walk. positional aNalysis Of VAriance (posNOVA) is a type of Analysis of Variance (ANOVA) that supports multifactor analyses of trajectory data with random walk. It is suited for studies where the participants have the complete freedom to roam in undefined manners, and that data mining of sub-trajectories is not feasible. eyeNOVA is a further enhancement of posNOVA to deal with gaze data trajectory data collected from devices such as Microsoft HoloLens v2. Unlike traditional eye trackers which produce 2D trajectories, these devices produce rays which require additional pre-processing.

In this chapter, we provide the background information on trajectory analysis. Then, we describe posNOVA and its procedure. Then, we discuss how to perform eyeNOVA–a variant of posNOVA for 3D gaze data from hardware like Microsoft HoloLens v2. To support the New Statistics initiative which encourages the visualization of estimates and their confidence interval, we also discuss deriving the confidence interval for mean squared displacement (MSD), the type of index number used to describe a trajectory.

A.1 Background Information

This section provides the technical background necessary to understand eyeNOVA. It includes a brief introduction to trajectory data, indexing trajectory, and gaze data analysis.

A.1.1 Types of Trajectory Indexing

We can categorize trajectories as two main types: one with obvious targets, and one without. Where there are obvious landmarks, we can compute a tortuosity index such as ST or the straightness value [Almeida et al., 2010]. The index indicates how much a participant has deviated from the straight and the shortest paths from two points. However, without obvious landmarks and with the participant being able to freely move, we cannot use tortuosity. Instead, we analyze how much movement has occurred within a space. The intensity of use is one such measure; it is a ratio of L, the path length, and $\sqrt{A}$ where A is the area [Almeida et al., 2010]. While intuitive, this measure is restricted to 2D spaces. Work in HCI, on the other hand, can involve movement in 3D space. For instance, our synoptic study allows the participants to stand up and walk around the glyph fields. MSD is a superior alternative because it is usable in both 2D [Almeida et al., 2010] and 3D spaces [Poupard et al., 2019].

A.1.2 Random Walk

Random walk describes random directions that the participant can take in the study. It is important to note that a random walk can involve any type of movement, not just walking. For instance, in a study where we measure the participants' gaze trajectories, the unpredicted gaze movement constitutes random walks. There are multiple degrees of random walk. At the most basic level, the participants' movement can be completely described using a random function [Codling et al., 2008]. However, many times, there are still certain degrees of patterns. For example, in our synoptic study, while we cannot predict the next direction that the participants will gaze at, we can still find areas where all participants would generally focus on.

The level of experimental control in a study can affect the amount of random walk. A highly controlled study can have less random walk. For example, in Hu et al. [2021], we asked the participants to search for out-of-view targets in virtual reality (VR). However, since the participants had guidance from the visual cueing systems, they did not deviate much from the intended trajectories. This results in a small degree of random walk. Meanwhile, our synoptic study allowed the participants to freely roam the glyph fields. This introduced more unpredictability and therefore, more random walk.

Mean Squared Displacement

Originally, MSD is a concept in physics. It represents the distance of a particle from its original point in a diffusion process [Balakrishnan, 2021]. However, it is also adopted for general trajectory analysis [Almeida et al., 2010]. Almeida et al. [2010] warn that MSD does not represent tortuosity; instead, it represents how "spread out" the participants are in a space.

There are multiple formulae for mean-squared displacement. In physics, the formulae assume that the trajectory is continuous–i.e., there is an infinitesimal amount from one point to another. As such, they are described as integral functions (e.g. Balakrishnan [2021]). However, for HCI, the trajectory data are sampled in a discrete and non-infinitesimal interval. This means we can use simpler formulae.

Assuming that the trajectory starts at $\vec{0}$, the discrete MSD formulae are:

- 2D version [Almeida et al., 2010]: $MSD = Var(X) + Var(Y)$
- 3D version [Poupard et al., 2019]: $MSD = Var(X) + Var(Y) + Var(Z)$

where Var is the variance function, X, Y, Z are sets of x-, y-, and z-positions within the trajectory. If the trajectory does not start at $\vec{0}$, an offset must be introduced. For instance, if the offset is $(1, 1, 1)$, the 3D formula becomes $MSD = Var(X - 1) + Var(Y - 1) + Var(Z - 1)$.

The distribution for MSD is unknown. However, observations from researchers indicate that, in general, a MSD distribution probably follows a log-normal distribution [Hoffman et al., 2006, Grebenkov, 2011]. Grebenkov [2011] thinks that if the MSDs come from a Gaussian process, then they follow a Γ-distribution. These distributions can possess extreme skewness with certain parameters.

A.2 Procedure

Here, we outline the general procedure for posNOVA. Before we can apply the test, we must first collect positional data. These can be a set of 2D and 3D positions. Then, for each trajectory, we compute its MSD using the formula outlined in Section A.1.2. It is important to point out that, if possible, we should also data mine the trajectories and perform exploratory data analysis. While the data in the synoptic study contains too much random walk for data mining to be effective, in a more controlled study, we can imagine segmenting trajectories into smaller sub-trajectories for closer examinations. Furthermore, clustering of positions or a heatmap analysis can be useful to identify areas where the participants moved to the most. If video data are available, using software like BORIS [Freiwald et al., 2018] to identify patterns before further analysis can also be useful.

A.2.1 General Hypothesis

The general null hypothesis for posNOVS is that all MSDs are equal regardless of the conditions. The alternative hypothesis is that at least one condition has a different MSD. Therefore, the post hoc test has the following hypotheses:

- $H_0 : MSD_a = MSD_b$
- $H_1 : MSD_a \neq MSD_b$

A.2.2 Modelling

Modelling a MSD distribution can be extremely difficult, because a general distribution of MSD is unknown [Hoffman et al., 2006, Grebenkov, 2011]. Therefore, we must rely on a nonparametric or a semi-parametric test. For the HCI community, Wobbrock et al. [2011], Elkin et al. [2021] suggest using ART-ANOVA to test data with unknown distributions. However, in our experience and from observations [Hoffman et al., 2006, Grebenkov, 2011], MSD distributions can be log-normally distributed. If MSDs are log-normally distributed, then they are highly skewed Durivage [2022] and are incompatible with ART-ANOVA [Luepsen, 2017]. Applying a log-transformation can alleviate the issue of extreme skewness.

An alternative to ART-ANOVA is PERMANOVA. PERMANOVA is a nonparametric test developed by Anderson [2017]. Unlike ART-ANOVA and other nonparametric tests (e.g., Kruskal-Wallis, and Friedman's Test), it incorporates the use of simulation with data permutation. Furthermore, it can support multivariate dependent variables, and continuous independent variables. A second alternative is creating a semi-parametric GLM. A third alternative is to apply log-transformation and proceed with the normal parametric methods. Log-transformation is used by and suggested by multiple works; Poupard et al. [2019] applied it directly on MSDs, and Bailey et al. [2022] recommend applying log-transformation to the raw data before computing MSDs.

A.3 eyeNOVA

eye aNalysis Of VAriance (eyeNOVA) is an extension of posNOVA to support 3D gaze analysis. It is essentially a pre-processing procedure plus posNOVA. If the gaze data are already in 2D, we can proceed with posNOVA. However, 3D gaze ray data from devices like Microsoft HoloLens require additional steps before posNOVA can be applied.

A.3.1 Define the 2D Surface

To perform eyeNOVA, the researcher must convert the gaze directions to 2D points of intersections (PoI). The surface should represent where the participants were focusing. For instance, in the synoptic study, the participants were asked to focus on the glyph field which we mathematically define as a horizontal plane with the tablet's centre as the origin. We can then compute the points of intersections between the gaze direction ray with the plane.

Computing the points of intersection can be a time-consuming process. Therefore, before computing the points of intersections, we should try to identify rays that obviously will not intersect the surface. For instance, if the surface is below the participants' eyes, we excluded the gaze direction when the participants' pitch angles were above 0^o because these directions would never intersect the gaze plane.

A.3.2 Trimming

If the 2D surface has a defined boundary, we can simply exclude the gaze intersections that are outside the boundary. Otherwise, we need to use techniques like clustering to define where to trim the data. An example of this technique is `tclust` by Fritz et al. [2012].

A.3.3 Proceeding to posNOVA

Once the pre-processing is complete, the researcher can perform posNOVA on the transformed data.

A.4 Descriptive Statistics

Test statistics should be complemented by descriptive statistics. Choosing the right descriptive statistics (i.e. mean, median, or mode) and the measure of spread (e.g. variance, standard deviation) can aid the understanding of the hypothesis testing.

For future work, we propose a new type of estimator for MSD representing the population value of MSDs of multiple trials within the same condition. This estimator is based on the pooled variance formula (adopted from Zientek and Yetkiner [2010]):

$$MSD_{pooled} = \frac{\sum_{t=1}^{T}(n_t - 1)MSD_t}{\sum_{t=1}^{T}(n_t - 1)}$$

where t is a trajectory index, T is the number of trajectories, n_t is the number of points used to compute MSD_t. The pooled MSD is the estimate for the actual MSD for the condition. The confidence interval for the estimator is as follows (adapted from [Ghilani, 2018]):

$$\left[\frac{\nu MSD_{pooled}}{\chi^2_{\nu,0.5\alpha}}, \frac{\nu MSD_{pooled}}{\chi^2_{\nu,1-0.5\alpha}}\right]$$

where ν is the denominator of the MSD_{pooled}. MSD_{pooled} has been used in prior work. For example, Black et al. [1982] compared MSD_{pooled} in their human posture study.

We must caution that the confidence interval of MSD_{pooled} is not the same as the sample distribution of MSD. Therefore, while the confidence interval of MSD_{pooled} follows a χ^2-distribution, the sample distribution can be quite different (e.g., Γ-distribution, and a log-normal distribution).

We note that we did not compute MSD_{pooled} for our trajectory analyses in the synoptic study (Chapter 5). This is because this type of estimate is simply a proposal at this point. More research is necessary to actually justify its use. Furthermore, we reported mean MSDs and their standard deviations instead, because the reader could use them, through the method of moments, to reconstruct the Γ-distribution, a type of distribution used in the ANOVA analysis.

A.5 Visualization

Since some HCI practitioners adopt the New Statistics introduced by Cumming [2014], visualization is now playing a very important role in results representation. Therefore, we provide some recommendations on how to supplement posNOVA, and eyeNOVA with visualization.

A.5.1 Heatmap

Heatmap is a popular visualization method used in gaze analyses [Othman et al., 2020]. A heatmap showing the areas visited by the participants is helpful for comparison among the conditions. While it does not directly communicate the posNOVA/eyeNOVA statistics, we can still expect conditions with higher MSDs to produce more spread-out heatmaps and conditions with lower MSDs to produce more concentrated ones. An example of this is Fig. 5.10 in Chapter 5.

A.5.2 Histogram and Shape Visualization Techniques

We recommend the histogram and other related techniques (e.g. the violin plot) can serve as techniques for exploratory data analysis, because they can help determine the appropriate type of statistical test. For instance, if histograms show the MSD distributions to be highly skewed, then ART-ANOVA should be avoided. However, presenting the visualizations to the readers must be done with some caution. If MSDs are highly skewed, their visualizations can be difficult to interpret by the readers.

A.5.3 Point-and-Whisker Plots

The point-and-whisker plot is a method of presenting an estimator and its confidence interval. Assuming that the data are normally distributed, the estimator is the mean and the confidence interval is constructed by inverting the t-test or the Z-test.

MSDs, on the other hand, are not normally distributed. Since they are variance, their confidence intervals follow χ^2-distributions. To construct a point-and-whisker plot for MSDs, the estimator is the pooled MSD and the confidence interval is calculated from the confidence interval formula for the pooled MSD. Another alternative is bootstrapping. Due to the potential skewness in the MSD distribution, we should avoid using the Hodges-Lehmann method [Hodges Jr. and Lehmann, 1963] method which essentially involves inverting the Wilcoxon signed-rank test [Rosenkranz, 2010].

Eyeballing a Point-and-Whisker Plot: A Warning

Cumming [2014] states that we can guess if two means are different or not, based on how close the points-and-whisker plots are. If the whiskers are far apart and do not overlap, the means are likely to be statistically different from each other. This intuition does not apply to MSDs, because MSDs are variances. To compare two sets of between-subject MSDs, we apply the F-test which involves comparing how many times one pooled MSD is larger than the other one. For within-subject variances, we apply the Pitman-Morgan test to see how well the MSDs are correlated to each other. The Pitman-Morgan involves testing if Pearson's r correlation of two variables U, V is zero or not [Yamaguchi et al., 2023]. $U = X_{pre} + X_{post}$ and $V = X_{post} - X_{pre}$, and X_{pre}, X_{post} are the data from the pre-test and the post-test.

We must be cautious when using the F-test as a test for between-subject MSDs, because the F-test is not very robust [Hosken et al., 2018]. The Pitman-Morgan test is also very senstive. While multiple alternatives to the Pitman-Morgan test exist, they can be impractical to apply and/or still provide inaccurate results [Yamaguchi et al., 2023]. Additionally, the Pitman-Morgan test and its variants require us to solve for the "data" that have given rise to MSD_{pooled}. These "data" are not the same as the raw trajectory data. Rather, they are the solution vectors $(\mathbb{X}_{pre}, \mathbb{X}_{post})$ to the following equations:

$$MSD_{pooled,pre} = \mathbb{E}(\mathbb{X}_{pre}^2) - (\mathbb{E}(\mathbb{X}_{pre}))^2$$
$$MSD_{pooled,post} = \mathbb{E}(\mathbb{X}_{post}^2) - (\mathbb{E}(\mathbb{X}_{post}))^2$$

The equations can be further simplified by setting $\mathbb{E}(\mathbb{X}_{pre}) := 0$ and $\mathbb{E}(\mathbb{X}_{post}) := 0$:

$$MSD_{pooled,pre} = \mathbb{E}(\mathbb{X}_{pre}^2) = \frac{1}{|\mathbb{X}_{pre}|} \sum_{i=1} \mathbb{X}_{pre}^2$$
$$MSD_{pooled,post} = \mathbb{E}(\mathbb{X}_{post}^2) = \frac{1}{|\mathbb{X}_{post}|} \sum_{i=1} \mathbb{X}_{post}^2$$

We believe that solving for $\mathbb{X}_{pre}, \mathbb{X}_{post}$ is impossible without the use of simulation (i.e.the Monte Carlo method). Even if we manage to find the solutions, the "data" may not meet the requirements for the Pitman-Morgan test or its variants. Therefore, we recommend nonparametric tests for between-subject data over these tests.

If we wish to visualize the confidence intervals of MSD_{pooled}, we must implement additional visual elements to prevent the reader from directly comparing the plots (e.g. drawing stars between statistically significant pairs of confidence intervals like in Fig. A.1). The F-test, the Pitman-Morgan test, and the variants of the Pitman-Morgan test have their own challenges. If an omnibus test has already been performed, we suggest using posthoc test results in lieu of these tests to indicate statistical significance.

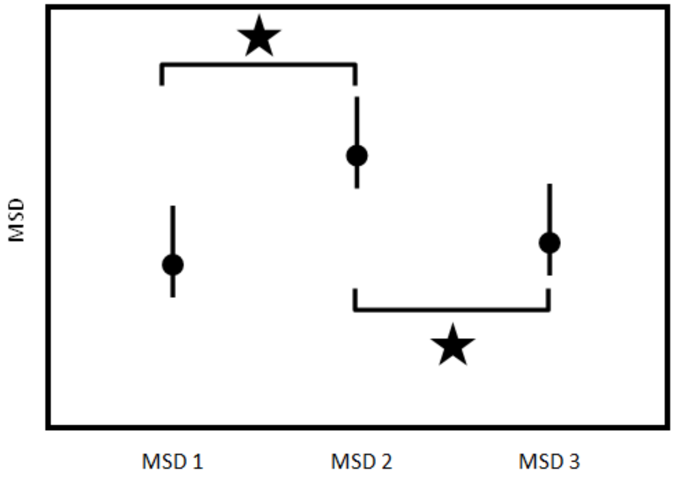


Figure A.1: Point-and-whisker plots of pooled MSD from a fictitious study. The whiskers represent sample confidence intervals. Since MSDs are not distributed like means, we cannot rely on "eyeballing" to determine their differences. Instead, we must rely on the stars between the plots.

Appendix B

Approval Letters from the Research Ethics Boards and Consent Forms

B.1 Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

This approval letter and the consent form in this section are for the Synoptic Study.

B.1.1 Approval Letter

See next page for the copy ▷▷▷

Social Sciences & Humanities Research Ethics Board
Amendment Approval

March 10, 2022

Hubert (Sathaporn) Hu
Computer Science\Computer Science

Dear Hubert (Sathaporn),

REB #: 2021-5726
Project Title: Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

The Social Sciences & Humanities Research Ethics Board has reviewed your amendment request and has approved this amendment request effective today, March 10, 2022.

Effective March 16, 2020: Notwithstanding this approval, any research conducted during the COVID-19 public health emergency must comply with federal and provincial public health advice as well as directives from Dalhousie University (and/or other facilities or jurisdictions where the research will occur) regarding preventing the spread of COVID-19.

Sincerely,

[SIGNATURE REDACTED]

Dr. Karen Foster, Chair

B.1.2 Consent Form

The participants received the form below via email before the study. When they came to the laboratory, we allowed them to review the physical copy of the form. They then signed the consent form at the beginning of the study using a web-based interface.

Form (Email, Physical)

See next page for the copy ▷▷▷

DALHOUSIE UNIVERSITY

CONSENT FORM

Project title: Designing a Tablet and Augmented Interface for Multiple Linear Regression

Lead researcher: Hubert (Sathaporn) Hu, Faculty of Computer Science, hs.hu@dal.ca

Other researchers
Derek Reilly, Faculty of Computer Science, reilly@cs.dal.ca
Ramanpreet Kaur, Faculty of Computer Science, rm216536@dal.ca
Hariprashanth Deivasigamani, Faculty of Computer Science, hr533370@dal.ca

Funding provided by: Mitacs Accelerate

Introduction
We invite you to take part in a research study being conducted by, Hubert (Sathaporn) Hu, who is a student at Dalhousie University. Choosing whether or not to take part in this research is entirely your choice. The information below tells you about what is involved in the research, what you will be asked to do and about any benefit, risk, inconvenience or discomfort that you might experience.

Please ask as many questions as you like. If you have questions later, please contact Mr. Hu.

Purpose and Outline of the Research Study
The purpose of the study is to evaluate a pair of interfaces for assessing multiple linear regression models. Our study will include the use of a tablet, an augmented reality (AR) headset as well as a digital tabletop interface.

Who Can Take Part in the Research Study
Participants in this study must have some knowledge in multiple linear regression. They must also be able to physically use the AR headset, the tablet, and the digital tabletop.

What You Will Be Asked to Do
You will use the tablet and the head-worn display device to create a multiple linear regression. At the same time, you will be asked to articulate the steps that you take. Afterwards, there will be a brief interview.

Possible Benefits and Risks
Benefits: Participation in this study does not provide a direct benefit to you. Instead, the results of the study will help us to fine-tune the interface before deploying for another study. You will

also receive cash reward for your participation even if you do not complete the study.

Risks: There is minimal risk in the study. While the AR headset may be uncomfortable to use, you will less likely have cybersickness than when using a virtual reality headset.

Compensation
You will receive $25 for your participant – even if you do not complete the study.

Management of Your Data
During the study, you will generate the following data:

- Interview and questionnaire data: During the interview and the questionnaires, we will record your responses using a computer, and an audio recorder.
- Software log data: The hardware devices have software that can log the actions that you perform. The AR headset can also video record what you see in the mixed reality; it never records your face.

Most of the data will be deposited into an open repository. We **will never** make the audio and the video recordings available publicly – only their transcriptions will be uploaded. All other data will be anonymized to ensure your privacy. When we are reporting the results of the study in a publication and in Hu's thesis, we will mostly report aggregated data (eg. average) and statistical models created from the data. We may quote you during the study. However, your name will not be attached to the quote.

If You Decide to Stop Participating
You are free to leave the study at any time. If you decide to stop participating during the study, your data will be automatically destroyed. After participating in the study, it will be impossible to destroy the data.

How to Obtain Results
The results of the study will be made available through Mr. Hu's thesis. Some of the results will be available through peer-reviewed publication. Some of the anonymized data may become available in a public repository. You can provide your email address again when consenting to the study to confirm that you would like notification when the results become available. Please keep in mind that you may have to wait for a significant period of time before receiving a notification, because publication process can be slow.

Questions
We are happy to talk with you about any questions or concerns you may have about your participation in this research study. Please contact Researcher Name (hs.hu@dal.ca) at any time with questions, comments, or concerns about the research study.

If you have any ethical concerns about your participation in this research, you may also contact Research Ethics, Dalhousie University at (902) 494-3423, or email: ethics@dal.ca (and reference REB file # 2021-5726).

Signature Page (Web-based)

Project Title: Designing a Tablet and Augmented Interface for Multiple Linear Regression

Lead Researcher: Hubert (Sathaporn) Hu, Faculty of Computer Science, hs.hu@dal.ca

I have read the explanation about this study. I have been given the opportunity to discuss it and my questions have been answered to my satisfaction. I agree to take part in this study. My participation is voluntary and I understand that I am free to withdraw from the study at any time.

_ I consent to be in this study.

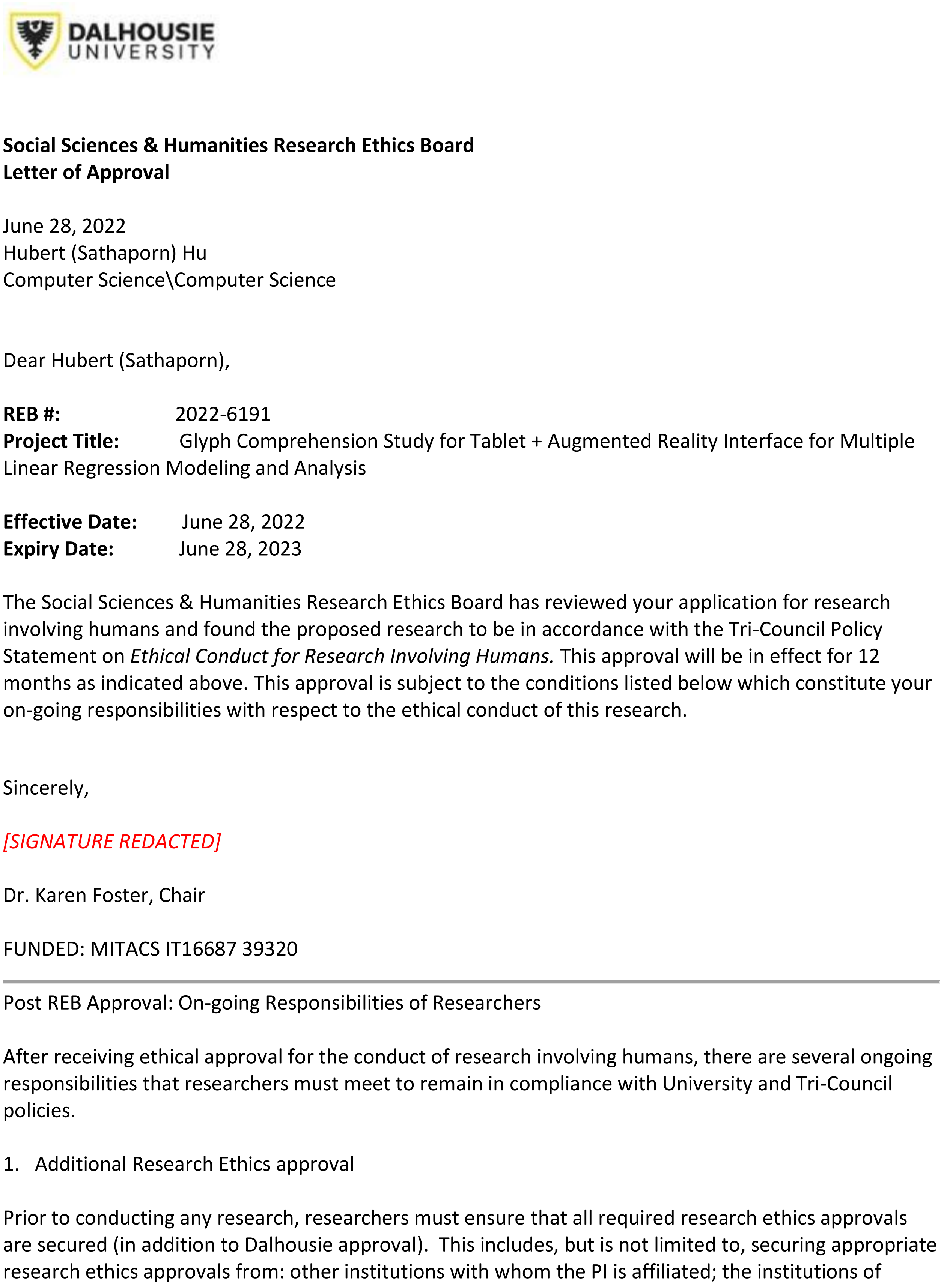
I would like access to pre-prints of publication related to this study. I also would like a link to the data on a public repository. Please notify me using this email address:

B.2 Glyph Comprehension Study Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

This approval letter and the consent form in this section are for the Elementary Study. The form is physical.

B.2.1 Approval Letter

See next page for the copy ▷▷▷



DALHOUSIE
UNIVERSITY

Social Sciences & Humanities Research Ethics Board
Letter of Approval

June 28, 2022
Hubert (Sathaporn) Hu
Computer Science\Computer Science

Dear Hubert (Sathaporn),

REB #: 2022-6191
Project Title: Glyph Comprehension Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

Effective Date: June 28, 2022
Expiry Date: June 28, 2023

The Social Sciences & Humanities Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans.* This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Sincerely,

[SIGNATURE REDACTED]

Dr. Karen Foster, Chair

FUNDED: MITACS IT16687 39320

Post REB Approval: On-going Responsibilities of Researchers

After receiving ethical approval for the conduct of research involving humans, there are several ongoing responsibilities that researchers must meet to remain in compliance with University and Tri-Council policies.

1. Additional Research Ethics approval

Prior to conducting any research, researchers must ensure that all required research ethics approvals are secured (in addition to Dalhousie approval). This includes, but is not limited to, securing appropriate research ethics approvals from: other institutions with whom the PI is affiliated; the institutions of research team members; the institution at which participants may be recruited or from which data may be collected; organizations or groups (e.g. school boards, Indigenous communities, correctional services, long-term care facilities, service agencies and community groups) and from any other responsible review body or bodies at the research site.

2. Reporting adverse events

Any significant adverse events experienced by research participants must be reported **in writing** to Research Ethics **within 24 hours** of their occurrence. Examples of what might be considered “significant” include: a negative physical reaction by a participant (e.g. fainting, nausea, unexpected pain, allergic reaction), an emotional breakdown of a participant during an interview, report by a participant of some sort of negative repercussion from their participation (e.g. reaction of spouse or employer) or complaint by a participant with respect to their participation, report of neglect or abuse of a child or adult in need of protection, or a privacy breach. The above list is indicative but not all-inclusive. The written report must include details of the situation and actions taken (or proposed) by the researcher in response to the incident.

3. Seeking approval for changes to research

Prior to implementing any changes to your research plan, whether to the risk assessment, methods, analysis, study instruments or recruitment/consent material, researchers must submit them to the Research Ethics Board for review and approval. This is done by completing the amendment request process (described on the website) and submitting an updated ethics submission that includes and explains the proposed changes. Please note that reviews are not conducted in August.

4. Continuing ethical review - annual reports

Research involving humans is subject to continuing REB review and oversight. REB approvals are valid for up to 12 months at a time (per the Tri-Council Policy Statement (TCPS) article 6.14). Prior to the REB approval expiry date, researchers may apply to extend REB approval by completing an Annual Report (available on the website). The report should be submitted 3 weeks in advance of the REB approval expiry date to allow time for REB review and to prevent a lapse of ethics approval for the research. Researchers should note that no research involving humans may be conducted in the absence of a valid ethical approval and that allowing REB approval to lapse is a violation of the University Scholarly Misconduct Policy, inconsistent with the TCPS and may result in the suspension of research and research funding, as required by the funding agency.

5. Final review - final reports

When the researcher is confident that all research-related interventions or interactions with participants have been completed (for prospective research) and/or that all data acquisition is complete, there will be no further access to participant records or collection of biological materials (for secondary use of information research), a Final Report (available on the website) must be submitted to Research Ethics. After review and acknowledgement of the Final Report, the Research Ethics file will be closed.

6. Retaining records in a secure manner

Researchers must ensure that records and data associated with their research are managed consistent with their approved research plans both during and after the project. Research information must be confidentially and securely retained and/or disposed of in such a manner as to comply with confidentiality provisions specified in the protocol and consent forms. This may involve destruction of the records, or continued arrangements for secure storage.

It is the researcher's responsibility to keep a copy of the REB approval letters. This can be important to demonstrate that research was undertaken with Board approval. Please note that the University will securely store your REB project file for 5 years after the REB approval end date at which point the file records may be permanently destroyed.

7. Current contact information and university affiliation

The lead researchers must inform the Research Ethics office of any changes to contact information for the PI (and supervisor, if appropriate), especially the electronic mail address, for the duration of the REB approval. The PI must inform Research Ethics if there is a termination or interruption of his or her affiliation with Dalhousie University.

8. Legal Counsel

The Principal Investigator agrees to comply with all legislative and regulatory requirements that apply to the project. The Principal Investigator agrees to notify the University Legal Counsel office in the event that he or she receives a notice of non-compliance, complaint or other proceeding relating to such requirements.

9. Supervision of students

Faculty must ensure that students conducting research under their supervision are aware of their responsibilities as described above and have adequate support to conduct their research in a safe and ethical manner.

B.2.2 Consent Form and Signature Form

See next page for the copy ▷ ▷ ▷

DALHOUSIE UNIVERSITY

CONSENT FORM

Project title: Glyph Comprehension Study for Tablet + Augmented Reality Interfaces for Multiple Linear Regression Modeling and Analysis
Lead researcher: Hubert (Sathaporn) Hu, Faculty of Computer Science, hs.hu@dal.ca

Other researchers
Dr. Derek Reilly, Faculty of Computer Science, reilly@cs.dal.ca
Mohammad Raza, Faculty of Computer Science, mh421497@dal.ca

Funding provided by: Mitacs Accelerate

Who Can Take Part in the Research Study?

To participate in the study, you must have a corrected to normal eyesight. It is OK if you wear glasses or wear contact lenses. You cannot have colour-vision deficiency or colourblindness. If you participated in the April/May 2022 version of the study, you cannot participate in this one.

Introduction

The purpose of the study is to evaluate glyph-based visualization techniques. A glyph is essentially a marker on a map. Our study will include the use of a tablet, an augmented reality (AR) headset as well as a virtual tabletop interface. The study should take about 1.5 hour to complete. However, it may run slightly over time based on multiple factors such as your prior experience with AR.

Consent

Consent to be in the study can be withdrawn at any time without penalty. This means that if you can quit the study at any time and still being compensated. Your data will be destroyed if you withdraw early. Otherwise, we will destroy the data.

Benefits

By participating in the study, you will receive $15 CAD in cash and you get to try HoloLens v2. You will also help with advancing the field of immersive analytics.

You should be aware that to receive the compensation, you must affirm that you will report the income to the Canadian Revenue Agency (CRA). You must provide your name and signature on a form to us. CRA requires us to keep this form for at least 7 years after your participation, and it can request us to show the form. Since the amount is small, it is unlikely that the CRA will ask us to show the form.

Risks

This study is a minimal-risk study. Therefore, you will not experience any bodily or psychological harm. Still, you may experience some discomfort due to the weight of HoloLens, its screen, and frequent turning. If you experience any discomfort, you should take a break.

What You Will Be Asked to Do

You will perform the following tasks:

- Complete an Ishihara test for colour-defiency and colourblindness. If you do not pass the test:
  - We will encourage you to see a specialist as we cannot officially diagnose you.
  - We will not allow you to proceed with the study.
- Complete a paper-based demographic background questionnaire.
- Complete a digital questionnaire while looking at 3D visualization in augmented reality.
- Complete a paper-based questionnaire that compares two visualization techniques.

Data Collection and Management

During the study, we will collect the following type of data:

- Paper data are collected from the questionnaires. These data will be digitized at the end of the data collection period, and we will destroy the physical copies.
- Video recording is from the HoloLens. The researcher will transfer the videos from the HoloLens into a server/computer that only the research team can access. The audio component will be uploaded to Microsoft Azure for transcription. The audio data will not be on Microsoft's server.
- Touch gesture data are deduced from game engine data and recorded on the HoloLens.

All digital data will be put into computers that only we can access. Afterwards, we will upload all data, except for audio data, into a public repository. All publicly available data will be anonymized and we will err on the side of removing too many data to preserve your anonymity. We may show some of the video recordings in public presentation without the audio components.

How to Obtain Results

Some of the results will eventually become publicly available as publications (eg. thesis, scientific papers). Providing your email address in the signature will allow us to send you the pre-prints of the publications. Publication process can be quite long; therefore, please do not expect to hear back from us soon.

Questions

We are happy to talk with you about any questions or concerns you may have about your participation in this research study. Please contact Hubert (Sathaporn) Hu (hs.hu@dal.ca) at any time with questions, comments, or concerns about the research study. If you have any ethical concerns about your participation in this research, you may also contact Research Ethics, Dalhousie University at (902) 494-3423, or email: ethics@dal.ca (and reference REB file # 20XX-XXXX).

Signature Page

Project Title: Glyph Comprehension Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis
Lead Researcher: Hubert (Sathaporn) Hu, Faculty of Computer Science, hs.hu@dal.ca

I have read the explanation about this study. I have been given the opportunity to discuss it and my questions have been answered to my satisfaction. I agree to take part in this study. My participation is voluntary and I understand that I am free to withdraw from the study at any time.

____________________________ ______________________________ ___________________
Name Signature Date

[Optional] Provide an email address to received pre-prints of Mr. Hu's manuscripts. It might take a long time before you hear back from us:

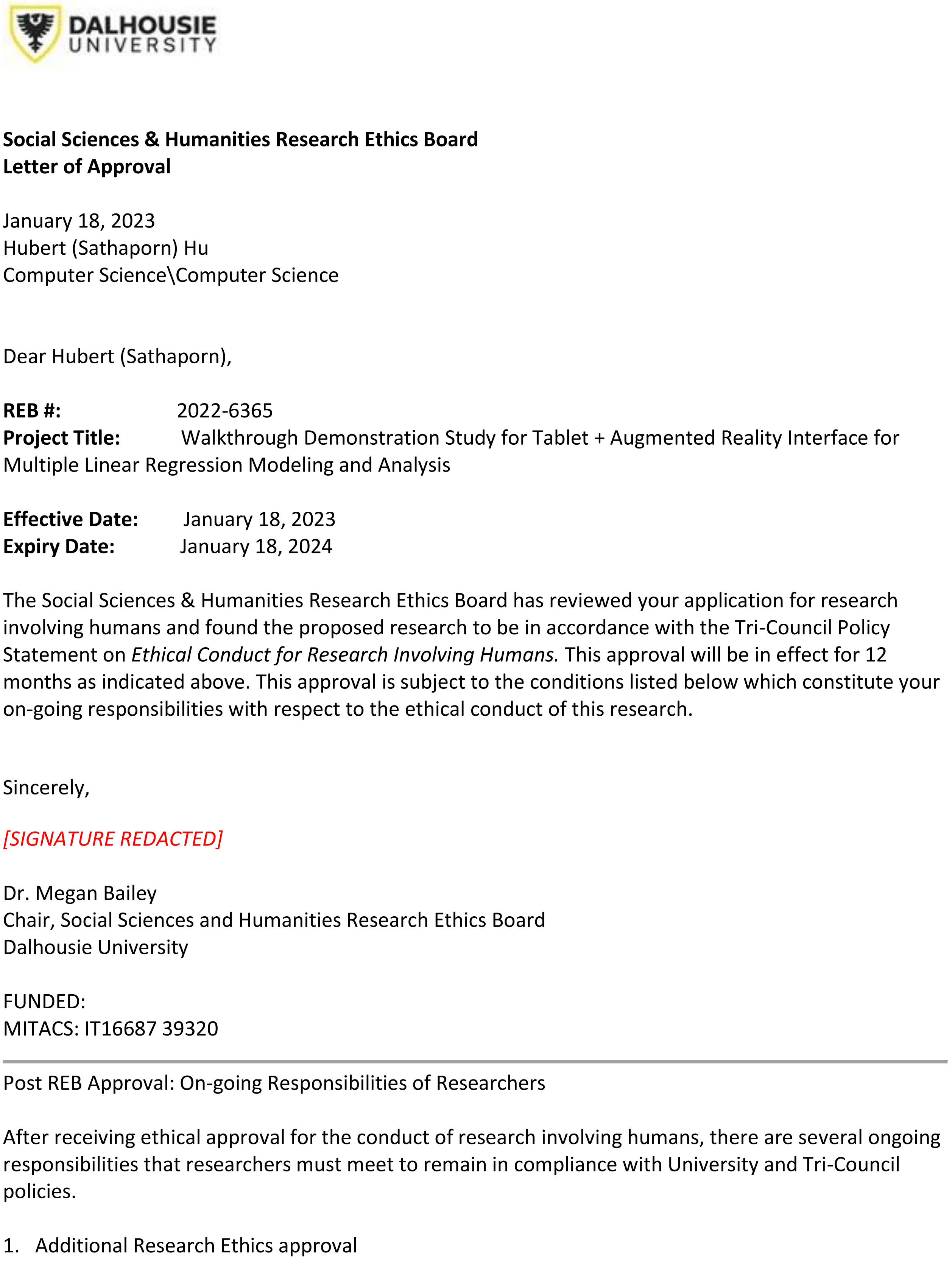
__________________________@_____________

B.3 Walkthrough Demonstration Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

Since we conducted this study at two locations, we applied for and received approval from two research boards (Dalhousie University, and Algoma University). These forms are for the Expert-feedback study.

B.3.1 Approval Letter (Dalhousie University)

See next page for the copy ▷▷▷



DALHOUSIE UNIVERSITY

Social Sciences & Humanities Research Ethics Board
Letter of Approval

January 18, 2023
Hubert (Sathaporn) Hu
Computer Science\Computer Science

Dear Hubert (Sathaporn),

REB #: 2022-6365
Project Title: Walkthrough Demonstration Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis

Effective Date: January 18, 2023
Expiry Date: January 18, 2024

The Social Sciences & Humanities Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans.* This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Sincerely,

[SIGNATURE REDACTED]

Dr. Megan Bailey
Chair, Social Sciences and Humanities Research Ethics Board
Dalhousie University

FUNDED:
MITACS: IT16687 39320

Post REB Approval: On-going Responsibilities of Researchers

After receiving ethical approval for the conduct of research involving humans, there are several ongoing responsibilities that researchers must meet to remain in compliance with University and Tri-Council policies.

1. Additional Research Ethics approval

Prior to conducting any research, researchers must ensure that all required research ethics approvals are secured (in addition to Dalhousie approval). This includes, but is not limited to, securing appropriate research ethics approvals from: other institutions with whom the PI is affiliated; the institutions of research team members; the institution at which participants may be recruited or from which data may be collected; organizations or groups (e.g. school boards, Indigenous communities, correctional services, long-term care facilities, service agencies and community groups) and from any other responsible review body or bodies at the research site.

2. Reporting adverse events

Any significant adverse events experienced by research participants must be reported **in writing** to Research Ethics **within 24 hours** of their occurrence. Examples of what might be considered "significant" include: a negative physical reaction by a participant (e.g. fainting, nausea, unexpected pain, allergic reaction), an emotional breakdown of a participant during an interview, report by a participant of some sort of negative repercussion from their participation (e.g. reaction of spouse or employer) or complaint by a participant with respect to their participation, report of neglect or abuse of a child or adult in need of protection, or a privacy breach. The above list is indicative but not all-inclusive. The written report must include details of the situation and actions taken (or proposed) by the researcher in response to the incident.

3. Seeking approval for changes to research

Prior to implementing any changes to your research plan, whether to the risk assessment, methods, analysis, study instruments or recruitment/consent material, researchers must submit them to the Research Ethics Board for review and approval. This is done by completing the amendment request process (described on the website) and submitting an updated ethics submission that includes and explains the proposed changes. Please note that reviews are not conducted in August.

4. Continuing ethical review - annual reports

Research involving humans is subject to continuing REB review and oversight. REB approvals are valid for up to 12 months at a time (per the Tri-Council Policy Statement (TCPS) article 6.14). Prior to the REB approval expiry date, researchers may apply to extend REB approval by completing an Annual Report (available on the website). The report should be submitted 3 weeks in advance of the REB approval expiry date to allow time for REB review and to prevent a lapse of ethics approval for the research. Researchers should note that no research involving humans may be conducted in the absence of a valid ethical approval and that allowing REB approval to lapse is a violation of the University Scholarly Misconduct Policy, inconsistent with the TCPS and may result in the suspension of research and research funding, as required by the funding agency.

5. Final review - final reports

When the researcher is confident that all research-related interventions or interactions with participants have been completed (for prospective research) and/or that all data acquisition is complete, there will be no further access to participant records or collection of biological materials (for secondary use of information research), a Final Report (available on the website) must be submitted to Research Ethics. After review and acknowledgement of the Final Report, the Research Ethics file will be closed.

6. Retaining records in a secure manner

Researchers must ensure that records and data associated with their research are managed consistent with their approved research plans both during and after the project. Research information must be confidentially and securely retained and/or disposed of in such a manner as to comply with confidentiality provisions specified in the protocol and consent forms. This may involve destruction of the records, or continued arrangements for secure storage.

It is the researcher's responsibility to keep a copy of the REB approval letters. This can be important to demonstrate that research was undertaken with Board approval. Please note that the University will securely store your REB project file for 5 years after the REB approval end date at which point the file records may be permanently destroyed.

7. Current contact information and university affiliation

The lead researchers must inform the Research Ethics office of any changes to contact information for the PI (and supervisor, if appropriate), especially the electronic mail address, for the duration of the REB approval. The PI must inform Research Ethics if there is a termination or interruption of their affiliation with Dalhousie University.

8. Legal Counsel

The Principal Investigator agrees to comply with all legislative and regulatory requirements that apply to the project. The Principal Investigator agrees to notify the University Legal Counsel office in the event that they receive a notice of non-compliance, complaint or other proceeding relating to such requirements.

9. Supervision of students

Faculty must ensure that students conducting research under their supervision are aware of their responsibilities as described above and have adequate support to conduct their research in a safe and ethical manner.

B.3.2 Approval Letter (Algoma University)

See next page for the copy ▷▷▷

Algoma
UNIVERSITY

RESEARCH ETHICS COMMITTEE
CERTIFICATE OF APPROVAL

FILE NO: 032-202223

PROJECT NAME: Walkthrough Demonstration Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis.

PRINCIPAL RESEARCHER: Hubert Hu

*has been considered by the Ethics Committee and is **APPROVED***

ETHICS Approval date: March 12, 2023 **ETHICS Expiry date: March 12, 2024**

It is the Principal Researcher's responsibility to ensure that all researchers associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher must notify the REB Chair, via amendment or progress report, of…

- Any significant change to the project; reasons for that change, highlighting ethical implications (if any);
- Serious adverse effects on participants and action(s) taken to address those effects;
- Any other unforeseen events or unexpected developments that merit

notification; ▪ Any change in Principal Researcher;

- A delay of more than 12 months in the commencement of the project;

and, ▪ Termination or closure of the project and the reasons for this.

Additionally, the Principal Researcher is required to submit…

- An Annual Report every 12 months for the duration of the project;
- A Request for Extension of the project prior to the expiry date, if applicable;

and, ▪ A detailed Final Report at the conclusion of the project.

The Ethics Committee may conduct an audit at any time.

SPECIAL CONDITIONS: NONE

SIGNED: *[SIGNATURE REDACTED]*

DATE: **March 12, 2023**

(REB Chair)

B.3.3 Consent Form and Signature Form

See next page for the copy ▷▷▷

DALHOUSIE UNIVERSITY

CONSENT FORM

Algoma UNIVERSITY

Project title: Walkthrough Demonstration Study for Tablet + Augmented Reality Interface for Multiple Linear Regression Modeling and Analysis
Lead researcher: Hubert (Sathaporn) Hu[1], Faculty of Computer Science, Dalhousie University, hs.hu@dal.ca

Other researchers
Derek Reilly, Faculty of Computer Science, Dalhousie University, reilly@cs.dal.ca
Mohammad Raza, Faculty of Computer Science, Dalhousie University, mh421497@dal.ca

Letter of Approval Numbers: REB2022-6365 (Dalhousie University), 032-202223 (Algoma University)

Funding provided by: MITACS Accelerate

Introduction

The purpose of the study is to evaluate an interface for geospatial data analysis in relation to current practices. This interface combines augmented reality with a tablet.

Consent

You can consent to be in the study by simply replying to the email that this file is attached to. Consent to be in the study can be withdrawn at any time.

When you consent in the first session, you also consent for all subsequent session(s). However, this does not prevent you from withdrawing later. Please let us know if you do not want to participate in the study anymore.

Benefits

The direct benefit is exposure to a novel technology that may complement your line of work. Furthermore, your participation will help advance the field of immersive analytics.

Risks

This study has small risks. There are things that you should keep in mind:

1. You will need to wear a Microsoft HoloLens v2, an augmented reality device during the second session. While this device is very comfortable when compared to other commercially available devices, you may experience some slight discomfort (eg. neck strain, eye strain, fatigue, minor headache), due to the weight of the device and its projection of images close to your eyes. If you experience slight discomfort, you can take

[1] Mr. Hu is also conducting the study at Algoma University.

breaks during the sessions. The device may also cause epilepsy in rare cases.

2. This study will ask you about your work. If you end up divulging sensitive information, you may put your employment at risk.

What You Will Be Asked to Do
The detail of the study is as following:

Session 1 (1 hour): In-person/Online

- Interviewing you for your work activities with geospatial data.
- Showing a video-based demo.
- Getting feedback.

Session 2 (1 hour): In-person [Optional]

- Demonstrating an AR prototype and demonstrate how this may help you with your research.
- Asking you for feedback on the prototype.
- **If you do not have time, this session is optional.**

Session 3 (1 hour): In-person/Online/Email-based [Optional]

- Showing an updated prototype.
- Asking you for feedback on the update.
- **If you do not have time, this session is optional or it can be email-based.**

The is a great degree of flexibility. Once consented to the study, we will discuss the sessions and how to schedule them.

Where Will the Study Be Conducted
Before the study begin, we will discuss the location of the study. If feasible, we would like to conduct all sessions at your workplace. However, if this is not possible, we will arrange to conduct Session 1 and 3 online. Session 2 must be in-person. It can happen at Dalhousie University or Algoma University.

Data Collection
We will collect the following types of data:

Physical Interview Notes: During the interview, we may record information on paper. If the interview note contains any relevant information, we will digitize it and store the digitized content on our lab's server. We will then destroy the physical copy.

Audio Recordings: If we interview you in-person, we will audio-tape you. We will upload the interview data to Microsoft Azure's transcription service for interview transcription. Since we will use a Canadian server for transcription, your data will never leave Canada.

If we interview you online, we will use Microsoft Team with our own Dalhousie University credentials. We will also use Microsoft Team's recording and transcription functions during the interview. During the interview, some of our data may be processed in the United States of America. As such, they are subjected to monitoring under the US Patriot Act. After the interview, the data will no longer be accessible to the US authorities. Instead, they will be securely stored on the Canadian soil.

Photos: We may collect screenshots of work instruments in Session 1 either photographed with a cellphone camera (in-person) or with a webcam (online). The work instruments may include the following: software screenshot, notes, and any tool relevant to your work.

Video Recordings: During Session 2, the HoloLens will video-tape your actions in the augmented reality. The HoloLens can only record what you see, and not your face. If you do not attend Session 2, this does not apply.

HoloLens/Tablet Log Data: During Session 2, the tablet and the HoloLens will record your body movement and touch gesture data. These data do not contain your name, and cannot be used to identify you in anyway. If you do not attend Session 2, this does not apply.

Publication of Collected Data

We will keep all of the data that have not been destroyed on our lab's server and on machines that are only accessible to the research team. These data will never be made public. However, with your consent, we:

(1) may quote you with your name replaced with a participation ID in publications and public presentations.
(2) may provide the screenshots of what you can see in the augmented reality, and the pictures of your work instruments in publication and public presentations.
(3) may provide a video snippet of what you can see through HoloLens in public presentations.

Confidentiality and Limit of Anonymization

We will attempt to keep your participation to this study anonymous. If we are allowed to quote you, we will replace your name with a participation ID. You do not need to worry about your face being visible in the screenshots and/or the videos in public presentations since the HoloLens does not record the face. Furthermore, we will never make audio data, including audio recording made by the HoloLens, publicly available.

Since we focus on recruiting experts, our participation pool is very small. This makes it somewhat easy for a reader to deduce that your workplace or laboratory has participated in the study from our publication. However, our publication will not any confirmation of specific workplace or laboratory members having participated in this study.

If you think that you have provided us with information that should not be publicly available, please let us know as soon as possible. This will allow us to remove this data from analysis publication. Once we submit a publication, it is already too late to remove this information. Please make sure to hide all sensitive information from your work instruments since we may take a screenshot of them. For instance, if the interface of your Geographical Information

System software is showing a content of a sensitive file before a session, you should switch to a different file or create a new blank file instead. If you do not feel comfortable sharing an instrument, please do not share it with us.

Withdrawing from the Study

You can withdraw from the study for any reason. There are two methods of withdrawal:

1. You can send us an email to request a withdrawal. The address is the same one for consenting.
2. You can verbally indicate so during one of the sessions.

After receiving a withdrawal request, we will destroy the data collected from you and we will not use them for the analysis. Once you have completed the study, we can only remove data that affect compromise your employment (eg. Data that the members of the public should not know) from the analysis. We will use the remaining data for analysis.

Please note that not participating in the optional sessions is not the same as withdrawing from the study. We will continue to use the data that we have collected.

How to Obtain Results

Some of the results will eventually become publicly available as publications (eg. thesis, scientific papers). If you would like to pre-prints of the publication, please let us know in the consent email.

Questions

We are happy to talk with you about any questions or concerns you may have about your participation in this research study. Please contact Hubert (Sathaporn) Hu (hs.hu@dal.ca) at any time with questions, comments, or concerns about the research study. If you have any ethical concerns about your participation in this research, please follow one of the procedures below:

- **Dalhousie University (Halifax-based Participant):** Please contact Research Ethics, Dalhousie University at (902) 494-3423, or email: ethics@dal.ca (and reference REB file # 2022-6365).
- **Algoma University (Sault Ste. Marie-based Participant):** Please file a Research Concerns & Complaint form available on https://algomau.ca/research/ethics-procedures/

Consent Procedure

To consent to the study, please answer the following questions by email. (The email address is hs.hu@dal.ca):

1. Are you aware while we will try to anonymize your as much as possible, someone may still be able to suspect that you have participated in the study through publications?
2. Are you aware that you should hide sensitive information before each session, and if you have accidentally leaked sensitive information, you must let us know as quickly as possible?
3. Do you allow us to anonymously quote you during in publications (eg. Academic papers and Mr. Hu's thesis) and in public presentations?

4. Do you allow us to provide screenshots of what you can see in the augmented reality during the walkthrough demonstrations in publications and in public presentations? As the HoloLens do not record your face, you do not need to worry about your face being visible.
5. Do you allow us to show video snippets of what you can see in the augmented reality during the walkthrough demonstrations in public presentations? The audio component will be removed and your face will not be visible in the video.
6. Do you consent to participate in the study?

Answering “Yes” to all questions means you consent to participate in the study. The reply email will be considered a documentation of your consent. Once we receive your consent, we will discuss the times and the locations of the study sessions.

If you would like us to send pre-prints of publication that describe the results of the study (eg. scientific papers, and Mr. Hu’s thesis), please let us know in the consent email as well.