12-2023

Using Statistical Clustering of Trajectory Data to Support Analysis of Subject Movement in a Virtual Environment

Martín Alejandro Galicia Avila
The University of Texas Rio Grande Valley

USING STATISTICAL CLUSTERING OF TRAJECTORY DATA TO SUPPORT ANALYSIS OF SUBJECT MOVEMENT IN A VIRTUAL ENVIRONMENT

A Thesis

by

MARTIN ALEJANDRO GALICIA AVILA

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Engineering Management

The University of Texas Rio Grande Valley

December 2023

USING STATISTICAL CLUSTERING OF TRAJECTORY DATA TO SUPPORT ANALYSIS OF SUBJECT MOVEMENT IN A VIRTUAL ENVIRONMENT

A Thesis
by
MARTIN ALEJANDRO GALICIA AVILA

COMMITTEE MEMBERS

Alley Butler, Ph.D.
Chair of Committee

Douglas Timmer, Ph.D.
Committee Member

Benjamin Peters, Ph.D.
Committee Member

December 2023

ABSTRACT

Galicia Avila, Martin A., Using Statistical Clustering of Trajectory Data to Support Analysis of Subject Movement in a Virtual Environment. Master of Science (MS), December, 2023, 76 pp., 12 tables, 15 figures, references, 38 titles.

The rising use of virtual reality software requires new methods of spatiotemporal analysis to enhance the user experience. This investigation is focused on forming clusters from spatiotemporal trajectories within a virtual environment and analyzing them in search of human immersion and behavior insights. The cluster centroids are plotted and analyzed to identify associations with the trajectories and user demographics.

DEDICATION

I dedicate this to my family. Martín Galicia Arratia, my dad and my best friend, who always had my back whenever I felt I could not keep going. You are a great example of how I can reach any level I commit myself to, we are a great team! My mom, Rosa Maria Avila Garcia who's hard working soul will always live inside me. As a kid I said I wanted to have a master's degree as you have, now I've finally done it! My sister, Iris Citlalli Galicia Avila who barely is joining this academic life. I am so proud of you and will always be.

Quiero dedicar esta tesis a mi familia. Martín Galicia Arratia, mi padre y mi mejor amigo, quien siempre tuvo mi espalda cuando sentía que ya no podía seguir. ¡Eres un gran ejemplo de cómo puedo llegar a cualquier nivel que me proponga, somos un gran equipo! Mi madre, Rosa María Ávila García, cuyo espíritu trabajador siempre vivirá dentro de mí. ¡De niño decía que quería tener una maestría, así como tú la tienes, y por fin lo logre! Mi hermana, Iris Citlalli Galicia Ávila quien apenas comienza esta vida académica. Estoy muy orgulloso de ti y siempre lo estaré.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Jianzhi Li and the I-DREAM4D Consortium for supporting this research when it began. I want to thank Dr. Sagnik Dakshit and Francisco Hinojosa Santillán who provided technical aid in some data analysis subjects where I lacked experience. Samuel Molina, Jorge Salazar and Jorge Lecea, thank you for your moral support while I completed this work. I want to also thank greatly UTRGV faculty Elizabeth Rodriguez, the custodians and police department whom I got on their last nerves by asking for office access on nights and holidays.

Words can't describe how grateful I feel for the support of Dr. Alley Butler and Dr. Douglas Timmer. Dr. Timmer taught me that no matter what life brings you, you should always be professional and responsible. His data analysis advice impacted greatly the quality of this investigation. Dr. Alley Butler is an example of a professional who seeks to help students to give their best work. I learned from Dr. Butler how to be a great graduate student who can complete his work, be engaged in other learning activities, and enjoy life at the same time.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Problem Description

The growing use of internet-of-things has made data mining a more relevant field of research. Mobile devices, now leave a digital footprint throughout their physical movement across space and these footprints help companies to find and understand new behavior patterns. These footprints consist of thousands of data points distributed across space and time, requiring the analysis and interpretation of vast amounts of data. Clustering algorithms are created to address this since they help analyze large datasets in a swifter way.

Gracia [1] performed an experiment where users took different trajectories in a virtual environment (VE) in response to 3 distractions. This experiment was also reported in Gracia et. al. [2,3]. The locations and orientations of the users' right and left hand, and head in a 3D space were recorded at intervals of 0.011 seconds. Gracia's thesis focused on examining the relationship between the user's performance and their gender. Although spatial temporal data was gathered, Gracia's work did not analyze the spatial temporal data for relationships to user performance. Using clustering algorithms, it is possible to analyze the trajectories taken in Gracia's experiment looking for similarities within user performance and user's demographics.

Background

The use of virtual reality (VR) technology has been growing due to the demand of its applications and the reduction in cost of mass production of its components. This increase in production and availability raises awareness in software design. VR companies seek to create more engaging environments while ensuring a safe use to be more competitive in the market. There are multiple areas in virtual reality that can benefit from these, e.g., medical, manufacturing, entertainment. All of them seek a design that helps the user to feel a strong presence to guarantee a strong performance in the environment, but the method of quantifying this data can be a challenge.

Gracia's Experiment

Numerous scientific studies have been conducted to calculate this "presence" or self-sense of being present. Gracia concluded an experiment where users in a VR path carried a plate on their non-dominant hand. This plate supported a ball which was balanced with the movement of the user. Several performance factors such as: track time, speed, times the ball is dropped, and how far the ball moved from the center of the plate, were used to compare with the user's self-perception given by a questionnaire.

Equipment Used. Tracking human motion requires tracking the degrees of freedom (DOF) that it involves. Hament [4] integrated movements in x, y, z, roll, pitch, and yaw to properly track motion in a VR snowboarding excercise. In the study, the user moved downhill and rotated to the sides, these are similar directions as those seen in Gracia's experiment [2] where the users rotated their hands in different directions such as down, up, and to the sides. To accurately capture the range of these movements, the 6 degrees of freedom are tracked with a HTC Vive system.

HTC Vive has two tracking base stations (known as 'lighthouses') that cover a 360-degree play area. They are placed in a designated area covering the 'path' that users follow. The user wears the head-mounted display and grabs the hand controllers during the virtual walk. The lighthouses are integrated with the sensors in the headset and controllers, allowing for a more accurate communication between them. When the lighthouse emits a light, the headset and controllers' sensors are activated, the lighthouse then measures this activation time. The lighthouses track the headset and controllers with the six DOF by measuring the position and orientation of the sensors when they are activated.

A study done by Supavich [5] proved that when the virtual environment has the presence of computer-generated social objects and has more 'cues' that resemble reality, the subjects show a higher level of social and spatial presence. Mel Slater [6] used logistic regression analysis to show that visual, auditory, and kinesthetic representation in a VE, associated with higher reporting of presence by users. Introducing the concept of 'stacking environments,' Slater included real environment factors to deepen the immersive level of the VE. The study mentions that when the user has a degree of association with its body in the VR, rather than seeing a cursor, there is a higher degree of unconscious presence. This environment factor combined with the quality of image, change in awareness, and field of view produce a deeper level of environment that influence the user's evaluation of the experience.

A Computer Science Team from UTRGV supported the development of the virtual environment (VE) used in Gracia's experiment [1]. This VE mainly comprised of:

- A 4 meters wide and 190 meters long path (shown in Figure 1.1)
- Surrounding trees and grass
- 3 visual and auditory distractions

- Visual representations of the hand controllers
- Visual representation of the 31 cm plate and 3 cm ball
- A coordinate system to identify motion

Figure 1.1
Virtual Path extracted from Gracia's study [1].

Given that the goals of Gracia's experiment are to study balance and coordination, the designed VE included 3 different stimuli to provide a change of environment to the user and record their spatial response. The stimuli are an explosion on the path shown in Figure 1.2, meteors falling from the left side, and a bird flock crossing through the path. These 'distractions' are not disclosed to the user previously.

Figure 1.2
Stimuli Events extracted from Gracia's study [1].

a) b) c)

Methodology. The realization of this experiment starts with a couple of forms that the user is required to complete. A consent form helps the user to acknowledge the equipment used, the purpose of the experiment, and the legal validation of it. A hand dominance survey is also completed which provides a good understanding of the configuration needed for the controllers; they are individually programmed to mimic real life hand-dominance. The arrengement ensures that the user balances the ball using his or her non-dominant hand.

Once the transition from real world to the VE happens by donning the equipment, a tutorial is played. This tutorial allows the user to become more familiar with the equipment. The tutorial takes place in a different VE than the actual test. At the end of this tutorial, the user is comfortable with the use of the controllers, balancing of the ball, and how to reset the ball in the plate in the situation that it falls.

For the experiment, the instructions given to the user is to walk through the marked trail or "path" as they balance the ball as close as possible to the center of the plate. The user navigates through the path as the 3 undisclosed obstacles appear. The level of immersion that these cause may force the user to move their hands and head, therefore changing trajectory to balance the ball. If the non-dominant hand moves in an angle where the ball falls, the user needs to reset the ball following tutorial instructions. This reset requires the click of a button on the controller to avoid movements that can alter the natural trajectory. The test ends once the user reaches the final destination. The virtual trajectories taken by all participants is recorded.

To ease the reintegration of the user to reality from the VE, the user is required to sit quietly during 10 to 15 minutes. This rest helped the human sensory system ensure that the human subject is back in the real life environment avoiding nausea, disorientation, or accidents due to changes in

environment. Once the user transitioned into the real environment, a 2 question survey was given where the recorded the level of "gamer" (videogame experience) is recorded and how difficult the subject considered it was to balance the ball in the VE.

Test and Results. Usoh [7] suggested to implement a questionnaire for VR subjects to answer using a scale from 1 to 7 to indicate their self-perception of presence in the environment. This scale is appropriate to quantify their "sense of being there" rather than analyzing open answers that may be subjective to the user. Gracia [1] employed a 7 question-survey for the users with the purpose of recording their level of self-perception during the experiment. The results from this questionnaire are statistically compared to the actual performance of the users during the test.

A digital file is created for each user in which the mentioned data is stored. The dataset is designed to organize the input into different variables such as track time, times the ball and plate are dropped along with the time when this occurred, the ball displacement from the plate, and the spatiotemporal positions of the right and left controllers as well as the headset.

To better analyze this collected data, the following statistical methods were used:

- Contingency Table
- T-Test
- Correlation
- Phase Plane

Contingency tables provided an organized structure to summarize and show data distribution among different groups or categorical variables. They helped identify patterns or relationships between the variables. Gracia [] performed 4 chi squared tests to compare performance during the experiment between two gender groups and analyzed any potential statistical difference.

Presence was divided in different categories. Table 1.1 and 1.2 have the contingency tables used. Table 1.1 shows 3 levels of presence (high, neutral, low) while Table 1.2 has 2 levels (high, not high). A null hypothesis was set to state that there is no statistical difference between male and females. The estimated expected value and the chi squared value were obtained following the formulas in Ott's book [8].

Table 1.1
X^2 results for presence extracted from Gracia's study [1]

a) 3 level

	Female	Male
High Presence	0.391304	0.391304
Neutral Presence	0.8	0.8
Low Presence	0.5	0.5
	P Value	
	0.184279	

b) 2 level

	Female	Male
Not High	1.285714	1.285714
High	0.391304	0.391304
	P Value	
	0.06704	

The test where presence was divided in 2 levels showed a higher statistical difference compared to the 3 level [3]. Another observation is that in the 2 level classification, female users showed a higher presence. Since a value of 95% was chosen, neither of both p values were considered statistically significant; the null hypothesis cannot be rejected. This means that the chi squared test could not prove that there is a statistical difference in self-perceived presence between males and females.

Table 1.2
X^2 results for ball drop extracted from Gracia's study [1]

a) General

	Female	Male
Low Drops	1.470588	1.470588
Neutral Drops	2.131579	2.131579
High Drops	0.071429	0.071429
	P Value	
	0.025385	

b) Per Event

	Female	Male
No Drop	3	3
Low Drop	2.382353	2.382353
High Drop	0.236842	0.236842
	P Value	
	0.003628	

The other 2 tests were done to consider the amount of times the ball was dropped among male and female categories. The General table divided the amount of times the ball was dropped into 3 levels (low, neutral, and high amount of drops) and it considered the complete path. The other table separated the amount of ball drops into no drop, low amount, and high amount during any of the 3 stimuli events. In both of the events, male users dropped the ball on fewer occasions than the female users.

The results of the chi squared tests had a low p value which, in respect to the confidence level, showed that there is a statistical significance difference between males and females. The null hypothesis was rejected. This circumstance means that male subjects in this experiment performed better, or have a better balance and coordination.

The next statistical method used was a t-test. This test, contrary to chi squared, assumes normally distributed data and compares the means of continuous variables rather than categorical values. Seven analyses were conducted comparing performance between male and female users, where a p-value higher than 0.05 represents a lack of statistically significant difference.

Multiple variables were considered and tested as shown in Table 1.3. The results showed that the only variable statistically significant was the amount of ball dropped during the stimuli events. Compared to the chi squared test mentioned earlier, the t-test considered the total amount of times the ball is dropped rather than a categorical value of zero, low, and high. The p-value of 0.013 means that the null hypothesis is rejected; there is a statistical difference between the amount of times the ball is dropped by males than by females in this experiment.

Table 1.3
The p-Value for T-Tests of Different Variables extracted from Gracia's study [1]

	P-Value
Track Time [secs]	0.958
Reset Ball Time [secs]	0.325
Balancing Time [secs]	0.994
Ball Drop on the complete path [score]	0.172
Ball Drop during the events [score]	**0.013**
Av. Ball Speed [m/s]	0.125
Delta Ball Speed [m/s]	0.601

A correlation measures the strength of linear relation between two axes (normally x and y) [8]. If there is a strong correlation coefficient, it indicates that x value is an accurate predictor of y value. The contrary can be assumed for a weak coefficient. Gracia [3] set for 0.5 to be the minimum coefficient value to consider a significant positive correlation and -0.5 for the maximum value to consider a negative correlation.

The performance variables used in the t-tests were analyzed using correlation, they were compared to the results of the 3 surveys: presence, gamer, and the balancing difficulty. The results were that neither the self-perception of presence or the videogames experience have a correlation with the performance.

The only correlations found were the performance of ball dropped and average ball speed compared to the difficulty survey within the female users. The female correlation coefficient between how difficult they found the experiment and the amount of times they dropped the ball was 0.623 compared to the 0.200 of males. The correlation coefficient between the difficulty survey and the average ball speed during the experiment was 0.588, while the male coefficient was 0.245. Based on these results, it may be concluded that the female subjects in this experiment had

a better interpretation of control in the VE, since they were more accurate in understanding their level of performance.

The phase planes are able to help the study of specific movements; however, for this experiment, the amount of total subjects made it hard for them to provide a deeper understanding into performance. Another method of analysis should be chosen for this analysis.

Further Research

Gracia's experiment gave favorable results into understanding the gap between presence and performance; however, the spatiotemporal data obtained about the users' trajectories was not used.

Studies using spatiotemporal data can help improve user experience by identifying areas where users may be struggling or where their perceptions are inaccurate. For example, if the research shows that users consistently overestimate their performance in a specific task, it may indicate that the virtual environment needs to provide more feedback on their low performance. Companies can use this information to improve the realism of the environment and make the experience less/more engaging or to keep users interested and motivated.

Another potential benefit that can be obtained from this data is an insight into safety of use. In the VR environment used, Gracia included 3 major distractions and recorded the user's response in their trajectories. This response can help assess how users move in reality while using a VR headset and prevent accidents with real world environment.

While there have been attempts at analyzing user's performance on a virtual environment, there is a lack of studies on the trajectories that these take. The user's position and movement add a layer of complexity to the data that can be challenging to analyze using traditional methods. Human balance and trajectories require the use of unsupervised machine learning since traditional methods have limits on spatiotemporal data analysis.

Research Objective

Given that there is remaining data in Gracia's experiment that was not utilized, the purpose of this work is to delve further into the analysis of the acquired data to determine if the subjects' demographics and perception of reality in the VE can influence their balance and trajectory. The general objective is to compare trajectories taken by the users' left/right hand and head by generating clusters that group them if similar. By comparing the trajectories to independent variables, a better understanding of human reaction to virtual stimuli can be achieved.

Through the use of clustering algorithms and data analysis techniques, this study investigates whether balance and motion response to 3 separate stimuli events in the VE correlate with the users' self-perception of performance and sense of presence.

Road Map

The structure of this thesis is made up by chapters. The first chapter summarizes the work done by Gracia de Luna [1] and provides an introduction to the relevance of clustering algorithms in data analysis of virtual reality datasets. The second chapter includes a literature review into different clustering techniques used for data analysis along with their benefits and which techniques work better with spatiotemporal data. Chapter three explains the methodology used to adapt de Gracia's data for clustering algorithms along with the use of these algorithms and their results. The fourth chapter dives into the results obtained from data analysis techniques. The final chapter concludes the study by connecting and interpreting the results from the analyses as well as including ideas for further research.

CHAPTER II

LITERATURE REVIEW

Big Data Analysis

The digital revolution has ignited an exponential growth in data generation and collection, this creates constant challenges for its analysis and interpretation. Big Data, the term given to the massive and complex collection of data, has become a critical component of modern research and industrial processes. Human activities such as online shopping, social media interactions, and even navigation patterns contribute immensely to this data collection. Algorithms about static and dynamic data are constantly being improved to understand this ever-growing data collection.

Zhu Y. [9] mentioned that Big Data "seeks to explore complex and evolving relationships among data." It is complicated to understand vast amounts of data. Wu [10] even compares it to blind men trying to size up a giant elephant that keeps growing. It is important for data analysts to develop strategies that can help read and understand the relationships among the data, even if this grows rapidly.

One kind of relationships that can be found in Big Data is preferences from online shoppers. Ozer [11] experimented with two common algorithms to analyze big data and concluded that online shoppers are more likely to be annoyed when they are either "bombarded" with information they do not need or frustrated when they receive less information than what they expected. This relationship between shopper' preferences can be key for the design of online pages that look to

increase their sales. Another type of patterns that big data can be used to identify is human emotions through virtual reality. Bulagang [12] coordinated an experiment with 20 subjects whose heart rate was tracked as they watched videos through a HTC Vive Virtual Reality headset. The results of this experiment showed that heart rate is a good potential approach in predicting human emotions.

Data Analysis and Virtual Reality

Patterns in virtual reality (VR) stand out as one of the novel sources of big data and it has been growing along with the mass production of virtual reality equipment. The data generated from VR environments has immense potential for understanding user behavior, preferences, and experiences. Health officials seek to understand how these environments may affect human psychology while companies seek to understand how humans can react to their designed virtual environments. Performance and self-evaluation of presence are two metrics commonly considered in these studies.

Hasenbein [13] performed an interesting experiment about social comparison behavior among students in a virtual reality classroom. The specific behavior examined is the participation of students by raising their hand to answer questions posed by the professor. The student's eyes were tracked; both the direction they were looking at and the pupil size when another student raised their hand. These eye movements were compared to a self-evaluation on class performance, and the results showed that students that were more aware of their peer's performance, gave themselves worse evaluations.

A research topic for virtual reality is which type of environments do users feel a greater sense of presence or in which the subject feels more "immersed." An experiment done by

Birenboim [14] consisted of participants cycling virtual routes with changing environmental characteristics and through a route with still, static images. A presence questionnaire was completed by participants to rate their experiences in the virtual routes. The results showed that participants ranked the immersive virtual environments higher than the still images.

Bouchard [15] performed an experiment that studied the subjective feeling of presence from individuals in a virtual reality environment through induced-anxiety. Bouchard analyzed the anxiety of two groups of participants with snake-phobia, which were exposed to the same environments but with a different state. The first group was told that their virtual environment was safe and contained no snake while the second group was led to believe that many dangerous snakes were lurking around. The results from this experiment was that presence, measured through a survey, was significantly higher in the group which believed snakes lurked around.

Another study where virtual reality performance and self-evaluation of presence were tracked is one done by Chuan [16]. He created virtual reality software that helps teach cognitive-motor needling skills needed for the performance of ultrasound-guided regional anaesthesia. The software achieved a perceived workload not significantly different from real world and also achieved acceptable levels of immersion, based on a survey completed by participants.

Pre-Process

The initial step before any data analysis, is to pre-process the data. Data is collected as events happen, showing location coordinates, qualitative or quantitative characteristics, and temporal position; this data can extend to be thousands of pages long. Therefore, it is helpful to "clean" or structure the data in a way that can be analyzed more easily. Xia [17] mentioned that it

is indispensable to clean the raw data before doing any analysis approach. The scope of the analysis might involve only certain characteristics from the data rather than all of them; cleaning, therefore, helps in reducing the time spent on irrelevant data points. For example, if the scope of the analysis is to look at half of the information from the database, it is best to delete the remaining half. Ansari [18] said that irrelevant attributes could negatively affect the similarity measures.

Apart from eliminating not useful data, the remaning information can be structured in a different format; titles could be added for easier location, dimensions can be zeroed to decrease computation time,

Trajectories

Big data is used to identify, as well, navigation patterns or trajectories. The mobility of smartphones, cars, and even particles in real life and virtual reality provide insights to traffic planning, video game design, and network coverage.

Due to common occurences of collisions between train and moose that cross a track, Peng [19] developed a finite element (FE) model of moose to investigate their motion trajectories. The team built models that represent moose using biological materials and, using simulation software, calculated how their path would impact high speed trains. The potential collision considered both intial position, and how and where the moose would land on the floor after a collision. Simulation was also applied to potential train derailment.

To improve volleyball serving training, Zhao [20] did a motion trajectory analysis of player's arms and joints. Zhao extracted body positions from video, or "gesture recognition," and calculated the bent arm angles, velocity, and even hitting force using mathematical formulas. The

results were proper trajectory analysis of serving techniques with feedback on how to improve performance.

Song [21] also extracted images to follow motion trajectories, but he used a different method for analysis. Song modeled a point tracking system to properly track vehicles on image sequences. Velocity curves and lane change were vehicle behaviors that Song analyzed. The result was a tracking system that alerted when vehicles changed lanes illegally or would potentially cause a traffic accident.

Spatiotemporal trajectories

Spatial data is information that can be found in a determined coordinate system, it comes along with qualitative and quantitative properties as Ansari [18] mentioned. A light turning on, a purchase made, and a sickness happening to a person are examples of events that can be found at a given position in space with qualitative and quantitative characteristics. Spatiotemporal data is the "motion" of data points which cause a pattern. That same light turning on in different places throughout time is an example of dynamic data since now, the pattern is taken into account.

Nanni M. [22] mentioned that data collection of transaction data and position has increased as more products are connected to each other through Intenet of Things (IoT). Mobile phones connect to each of the network antennas it passes by, generating a list of data that grows without limit. This data involves spatiotemporal coordinates that gives the exact position of the phone. The mining of this type of data is relevant to industries that seek to understand navigation paths.

Clustering

Spatiotemporal datasets represent new dimensions to track, thus, increasing the difficulty to analyze them. That is the reason that a common strategy to simplify this data is to cluster it. By

applying a similarity level to trajectories, it is possible to group them based on specific characteristics and therefore, reduce the amount of paths to analyze. There are multiple algorithms for trajectory clustering, each produces different results based on the data reviewed and the use that the analyst seeks.

There are several algorithms developed to analyze these spatiotemporal trajectories; Yuan [23] grouped into 5 main categories based on how they model data: model-based, hierarchical, grid-based, density-based, and partitioning methods. Hierarchical clustering requires large amount of computational time since it reads every data point as a cluster and then looks to merge it with others, Han [24]. Grid-based clustering, although used in many two dimensional path studies, do not work well for irregular paths. If the object in observation changes course by looping back to itself, zigzag, or other, it complicates the algorithm to properly cluster it. Model-based clustering works by looking at the spatiotemporal data and compare it to mathematical models to find the best match, Li [25]. Model-based analyses have the same limitation as grid-based since if the paths are irregular, they are less likely match with a mathematical model.

Density-based clustering is good for the detection of shapes and clusters from different sizes but they rely on other inputs, Agrawal [26]. The main two algorithms of this category are Density-Based Spatial Clustering and Ordering Points to Identify the Clustering Structure. These two algorithms are good for identifying nested clusters, or data points that might seem close in space, but far in other dimensions.

Density-Based Spatial Clustering (DBSCAN) functions based on the concept that regions with a high density of data points that are separated by sparser regions in the dataspace are identified as clusters. As Abdul [27] explained, the DBSCAN algorithm uses a radius input and neighborhood points to identify the core point and measure if its surrounded by the neighborhood

points in the radius, if so, these points become a part of one cluster. The core points that are in the radius of a point in that cluster are combined to it, making it bigger. The points near the cluster that are not core points, are added as well but the algorithm does not look for core points in their radius. When core points are detected far from the first cluster, they form a second cluster using the same methodology.

Birant [28] presented a spatial temporal version of DBSCAN that clustered weekly daytime and nighttime temperature records. By adding a radius for temporal distance, a "ST-DBSCAN" version was made. If the selected point has enough neighbors within the two radiuses, it is classified as a core object. This algorithm was used three times, to cluster similar sea surface temperature values, similar sea surface height residual values, and similar wave height. The run time in this experiment did not differ from the original DBSCAN runtime, stating that run time relies heavily on database complexity. Another use for ST-DBSCAN comes from Chimwayi. K. B. [29]. In this experiment, the algorithm was set to analyze a public health dataset. The data in use contained locations for West Nile Virus (WNV) traps and a temporal attribute which was the date of testing as well as other non-spatial attributes. To obtain the radius 1 (epsilon), a data ordering algorithm called k-nearest neighbor is used. The result of the ST-DBSCAN was the generations of clusters with positive and negative WNV cases but it lacked proper validation resources. The author stated that the larger the dataset, the more computational power it needed and at the time of writing the paper, there was no particular evaluation measure.

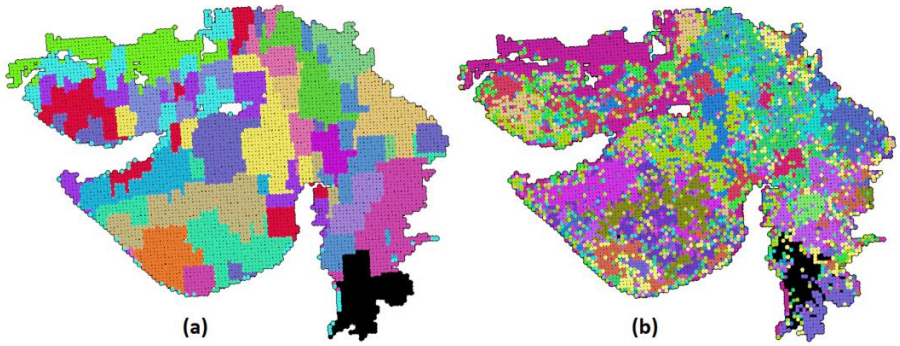
Ordering Points To Identify the Clustering Structure (OPTICS) has a clustering approach similar to DBSCAN, but contrary to it, OPTICS works better over clusters with different densities as Malhan stated [30]. This algorithm uses DBSCAN's input variables and also core distance and reachability distance. Core distance is the lowest possible radius to classify a point as core,

differing from the fixed radius in DBSCAN. The algorithm looks at a core point and looks for the minimum neighboring points, once the minimum is satisfied, the radius is calculated, this allows adjustment based on density levels. Reachability distance is used to order the data points based on the density region in which they are located, this helps the clustering step.

Agrawal [26] developed a version of OPTICS that generated spatial-temporal clusters of the presence of dense vegetation across Indian states. The algorithm steps for the ST-OPTICS are the same as Normal OPTICS but now considering two core distances (one more for time dimension). Figure 2.1, extracted from Agrawal's paper, shows the cluster results from ST DBSCAN and ST Optics. The difference between ST OPTICS and ST DBSCAN is evident in that (b) shows a clearer distribution among the clusters. Nested clusters are indeed differenced in ST OPTICS while ST DBSCAN's clusters are more "block-shaped."

Figure 2.1

Clusters made by: a) ST-DBSCAN; b) ST-OPTICS extracted from Agrawal's study [26]



Density-based clustering is a good tool for clustering spatial and spatiotemporal data, however there are some complications when using them to cluster whole trajectories. Due to their

nature of clustering scattered data points, the algorithms struggle when considering collection of data points (trajectories) as a group and then clustering them.

Partitioning methods work by separating data into independent groups based on distance from each other. The most famous algorithm in this category is k-means. This algorithm uses an input k value to find k groups among the data based on proximity. Since the best k might be unknown, it is necessary to repeat the program execution multiple times with different values until an optimal clustering is achieved. There are different versions of k-means, each used for special cases.

When looking at trajectories as paths from start to end, it is easier to identify them in a scatter plot rather than looking for portions of the path. ST-OPTICS and ST-DBSCAN are best for clustering scattered individual points or small portions of points, not complete paths; high computational times and RAM are required since they analyze every point. If the pre-process stage can group the full paths, k-means would help to cluster these full trajectories based on distance and save computational time.

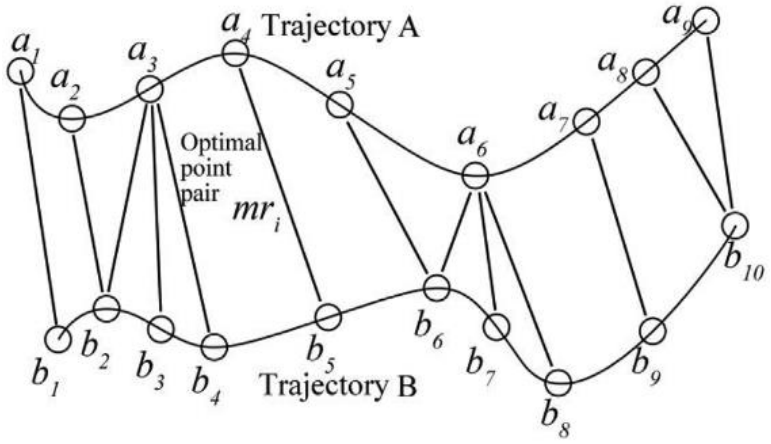
An important factor in the k-means algorithm is the metric that it uses to measure proximity since the measuring method has a high impact in the clusters formed. The most common distance metric is the Euclidean distance. This distance can be seen as the “straight” line between 2 points, different from Manhattan distance which cannot go directly but needs to travel in two dimensions; based on a grid layout. Measuring distances is essential for how clusters are found among spatiotemporal data.

For proper comparison between points from different trajectories, it is helpful to measure their “similiarity” or how far they are from each other in spatiotemporal coordinates. Zhuang [31] used Dynamic Time Warping (DTW) to measure similarity in aerial target’s trajectories. Chen

[32] explained that" DTW detects a 'warping' path through the matrix constructed by two sequences that minimize the cumulative distance" An advantage that DTW metric has is that it considers the sequence or track of past points that have existed prior to the most current when doing the similarity calculation. Zhuang [31] then used this similarity to feed a DBSCAN algorithm which performed the clusters. The results were proper clusters of dynamic enemy targets in a two dimensional which can indicate different tasks or tactics. Zhao [33] also used DTW and DBSCAN to form clusters. Trajectory A is made up of data points that are aligned to their corresponding "similar" point in trajectory B. Zhao provided Figure 2.2 to explain how DTW works.

Figure 2.2

Visual Representation of DTW alignment of two time-series provided by Zhao's study [33].



Dupas [34] used DTW to identify and align storm discharge time series of different lengths and with difference in phase. Once aligned, k-means clustering was used on them to identify common patterns in water quality in them. Another use for both techniques was to identify the behavior of pressure transients in water distribution systems as Xing [35] did. In this experiment,

DTW was used to measure the similarity between the pressure transients found by high-frequency pressure sensors and k-means then clustered them to discover the characteristic patterns.

Many experiments, like those from Dupas [34] and Xing [35], have been done using DTW to find similarity between trajectories and then use a clustering algorithm to generate the groups; however, they can also be combined. By using a clustering algorithm with DTW metric, clusters can be formed at a faster rate and with more distinguishable characteristics from one another.

Zhang [36] used K-means with DTW metric to cluster driving patterns in a x-y coordinate plane with time. Zhang managed to form clusters and assign them risk scores to vehicles approaching a signalized intersection based on their previous driving pattern. The risk score represented how likely a vehicle would exhibit risky driving in an intersection. Similarly, Chen [37] used DTW k-means clustering to separate lane-changing risk profiles into different categories. The lane-changing risk profiles exhibited the kinetic field generated by and on the vehicle; the higher the energy, the greater the risk indicator. Three categories were the result of this clustering: “uphill,” “bell,” and “downhill”. The Lane-changing risk profiles with “uphill” shape had the majority of the total risk profiles.

K-means with DTW metric was used by Jang [38] to cluster handwritings. The gesture recognition for a system functions by detecting the new handwriting gesture and comparing it to stored patterns to find the closest match. Jang proposed using this modified k-means algorithm to cluster the stored patterns so the new gesture would be compared to distinguishable groups rather than the complete list of patterns. The use of the algorithm decreased the number of reference patterns by 90% and increased the classification speed by 10-times of the normal speed.

As shown above, there have been studies that utilize algorithms to analyze virtual reality environments and clustering algorithms that facilitate trajectory analysis. However, there is a lack

of studies that combine both topics. Utilizing clustering algorithms to categorize virtual reality spatiotemporal trajectories for comparison with non-spatial metrics, is a nascent area of exploration.

CHAPTER III

PROCEDURE AND METHODOLOGY

Methodology Overview

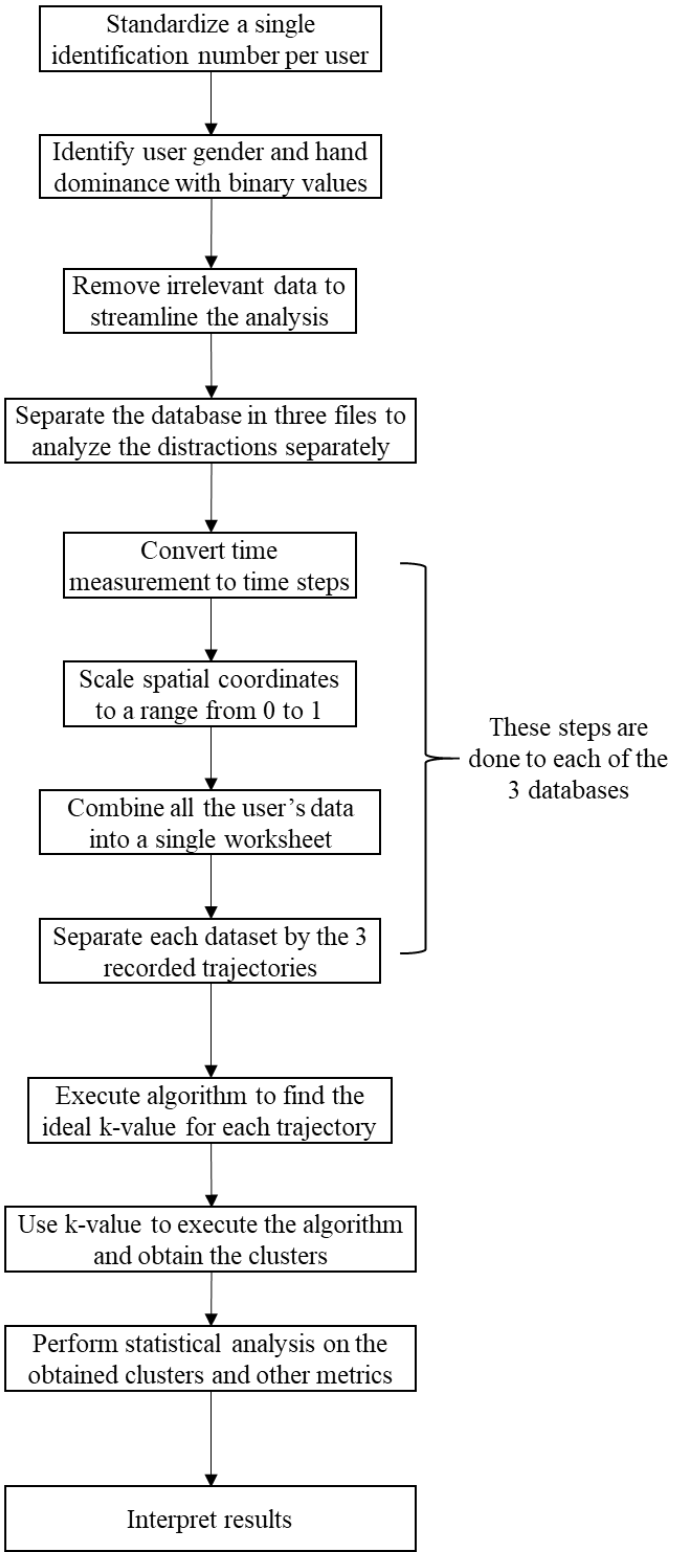
The objective of the experiment designed in this thesis is to perform statistical analysis of the collected data in order to determine if there are any significant interpretations to be drawn from the data. This objective requires a proper data structure that can be clearly interpreted by a proper algorithm to produce the desired results. There are several steps involved in obtaining the proper data structure and the desired results. These steps are summarized and shown in Figure 3.1.

Data Collection in Virtual Reality

The experiment conducted by Gracia [1] utilized a head-mounted display (HMD) and a pair of controllers, one for each hand. A tracking system, comprising of two position sensors, was also part of it. These sensors, important for the precision of the user's performance, were positioned at the start and end of the virtual path. The main purpose of the tracking system was to accurately track the six degrees-of-freedom (DOF) of the HMD and the hand controllers. The data from each user's session was recorded and saved into individual raw text files.

Figure 3.1

Flow Chart of Methodology Steps



Pre-Process Data

The raw database collected from the virtual reality set contains performance information from each user but in a structure that is not ideal for analysis. The 6 DOF: z,y,z roll, pitch, and yaw were recorded for head, dominant, and non-dominant hand across the spatiotemporal virtual space. Other data attributes are if the ball fell off the plate along the virtual path and the ball displacement from the center of the plate during each time interval.

The data was converted into an excel file, with a tab or worksheet for each user, by Gracia during his thesis methodology. This helps in reducing pre-process time.

Data Labeling

The first step into "cleaning" or re-ordering the given data was to assign the correct subject label to each worksheet and deleting the worksheets which do not have the same number of data points, since some users did not finish the virtual path. A column with the .txt file corresponding to the user is replaced with a column showing the subject identification number for easier identification when the tabs are combined. By ensuring each worksheet has the appropriate user identification, data misinterpretation is minimized.

Another step required to properly interpret the data is to label the independent variables of gender and hand dominance per user. Two columns are added to identify the hand dominance and gender. To make analysis more efficient, the user hand dominance and their gender are converted to the binary values 0 and 1. Left hand and Male are 0 while Right hand and Female are represented as 1. The columns with Left and Right Hand coordinates are changed to represent the appropriate dominant or non-dominant hand per user.

Remove Non-Essential Data

This paper focuses only on the user performance during the stimuli events, therefore, all time elapsed before and after each event is removed from the database. Three columns in the database indicate the exact time interval when the stimuli events (one column per event) started. The cell with "TRUE" that indicates the start of the event is the new beginning of the needed cells; all rows above this are deleted. The time range for the three distractions must be the same length to facilitate analysis. The stimuli event with the longest recorded duration is 'Meteors,' lasting 8 seconds or 720 time intervals. After the "TRUE" cell, 719 rows below are counted, and all rows starting from the 720^{th} cell downwards are deleted. By focusing solely on the time frame during the stimuli events, the analysis' attention is directed into the most crucial segments of the experiment, enhancing the precision of the results. If the trajectory of a user after a stimuli event is shorter than the 8-second time interval, then the user is removed from the database; this ensures that all trajectories have the same length.

To reduce computational time, the column with ball displacement and the columns with the spatial positions: roll, pitch, and yaw are removed from the database. All the worksheets in the database are selected, and deleting these columns from one worksheet simultaneously removes them all worksheets. Ball displacement is not considered in this research since its data represents another level of trajectory that occurs at simultaneously as users progress along the path. This analysis requires a different methodology from the one proposed to compare user's trajectories with the categorical metrics. Due to the nature of the roll, pitch, and yaw data as angles in space, the roll, pitch, and yaw data can provide negative numbers to show the turns and these are difficult to track when considering the user trajectory. Eliminating these parameters simplifies the dataset, allowing for a more refined analysis of user performance along the virtual path.

Stimuli Event Separation

The dataset is separated into 3 different versions (one for each stimuli event). There are three columns which indicate when each event starts, two of them are removed from all worksheets and saved as the database file for the remaining event. This same procedure is done for the two other stimuli events. The database separation helps each event have its own focus when analyzed and also reduces the computer loading time.

Scaled Coordinates

Temporal Standardization. The temporal position (time) is different for all events since it occurs gradually along the user's walk. This time is recorded in time intervals ranging from 0.01111 to 0.01114 of a second. The time intervals are transformed into time steps with a common starting time of zero and ending of 719. The Excel formula '`=IF(ROW()=2,0,INT((E2-E1+(F1/100))*100))`' is applied in a new column to the right of the time column. This formula verifies that the starting cell is assigned the value of zero. It then calculates the difference between the current and previous time intervals, scales the previous time step by dividing it by 100, sums the obtained difference to the scaled value, and then rounds down the result to the nearest whole number to eliminate any decimals.

Time steps represent discrete intervals of whole numbers, rather than continuous decimals, this offers benefits when analyzing the data. Reading the data uniformly in time steps is generally more convenient than calculating the exact decimals between positions. By measuring temporal position in time step intervals, algorithms can anticipate the structure of the data and use fixed-size memory blocks to reduce time and memory. Another benefit of this temporal transformation is easier visualization of graphs; plots with decimal numbers might make it hard to recognize trends

in the data, whereas, plots with whole numbers provide a clearer visual representation of trends over time.

Scaled Three Dimension Coordinates. A technique that helps interpret data more easily to scale all trajectories so their coordinates are located within the same spatial range, similar to the temporal standardization. Using a formula, the coordinates are scaled to a range between 0 and 1.

Database Organization

Merge Metrics and Spatiotemporal Coordinates. The next step of the data pre-process is to integrate the independent variables such as the survey answers, user gender, and user hand dominance into the location databases.

The users completed a questionnaire at the end of their virtual experience. The responses to the questions were given on a scale from 1 to 7, providing a quantifiable metric for analysis. The answers to these 8 questions are pulled from their table into the three databases by using the Excel XLOOKUP formula referencing their subject id. They are placed in the tab of their corresponding user as new columns. This action helps to improve the analysis using one single file.

Since all the user data is divided into different worksheets, the next step is to combine them into one. Using the VBA code shown in Appendix A: Worksheet Merge VBA, the worksheets are merged into a new worksheet called “Combined.” By merging the data, the analysis time can be reduced by eliminating time from identifying all tabs for each operation.

Data Separation by Situation. To reduce the algorithm’s loading time, the three databases are divided by trajectory. The result is 9 datasets with the same structure, whose data represents the spatiotemporal trajectories of 57 users and their independent variables, functioning metrics for comparison:

1. Explosion – Head
2. Explosion – Dominant Hand
3. Explosion – Non-Dominant Hand
4. Meteor – Head
5. Meteor – Dominant Hand
6. Meteor – Non-Dominant Hand
7. Birds – Head
8. Birds – Dominant Hand
9. Birds – Non-Dominant Hand

This data separation helps the algorithm to focus on each situation separately, reducing loading time and memory needed.

Algorithm: K-Means with DTW Metric

Elbow Method

K-Means was selected as the most appropriate algorithm to cluster the full trajectories that the users take in the virtual path. This algorithm requires for the initial input of "k" or the number of clusters desired to find in the data, and this approach is where the Elbow Method is useful. The elbow method technique consists of completing several k-means program executions using different k-values and plotting the resultant Within Cluster Sum of Squares (WCSS) whose formula 3.1 is shown below.

$$WCSS = \sum_{P_i \; in \; Cluster \; 1} distance(P_i, C_1)^2 + \cdots \tag{3.1}$$

where distance(P_i, C_1) is the distance from any point i to the centroid in cluster 1.

Formula 3.1 is repeated for the existence of more clusters and the result of each is summed to obtain the WCSS. The distance between points and their cluster centroid is squared to avoid negative numbers and to give more importance to the points furthest away from the centroid. The value of Within Cluster Sum of Squares shows how compactness of the data. The lower the score; the tighter the cluster.

When choosing the best k-value, a low WCSS is a good metric but not too low because that involves a high amount of clusters. The highest potential number of clusters equals the amount of data points. If this value is chosen as input, then the Within Cluster Sum of Squares is zero since each data point is its own centroid. A high WCSS is not good since it involves a low amount of clusters that do not properly categorize data points.

The Elbow Method plot shows the number of clusters against their obtained Within Cluster Sum of Squares. The "elbow" point in the plot indicates where the WCSS changes abruptly, showing the beginning of a trend that decreases slowly. This means that adding more clusters beyond this point gives little benefit to the results; the data points are already as close to the cluster centroids as they can meaningfully get.

The full (beginning to end) x,y, and z trajectories of the head, dominant hand, and non-dominant hand per user is then fed into the code (see Appendix A: Elbow Method Code) to consider them as positions in space and return their clusters. The code is executed using k-means integrated with dynamic time warping metric to obtain the "elbow" point of 9 scenarios: 3 stimuli events per point (head, dominant hand, non-dominant hand). A different k-value is expected to be obtained from each situation.

Forming Clusters

When the optimal k-values are identified for each of the nine scenarios, the focus now shifts to the clustering process. Each scenario, representing a different combination of stimuli events and trajectories is processed individually. This distinction is essential to capture the unique characteristics of user physical movements in response to each distraction event.

The execution of the k-means algorithm (see Appendix A: K-Means DTW Code), integrated with DTW metric, to these nine scenarios produces unique groups of similar movement patterns. The obtained clusters are expected to reveal variations in how users physically respond to the different distractions in the virtual environment. For example, the clusters created from the head trajectories following the explosion event might exhibit different characteristics compared to the clusters formed during the meteor event.

Jupyter Notebook environment is selected to execute the required code since it adapts best to running different code blocks instead of the full code. The code blocks calculate and show: cluster formation, color coded clusters in spatiotemporal space, trajectory distribution among clusters chart, and the statistical tests.

By examining the user distribution among the clusters across the three stimuli events, this research aims to explore deeper insights into user interactions within a virtual environment. The understanding of these clusters is useful in the subsequent analysis phase, where the relationship of clusters with other metrics, such as performance and presence of self-evaluation, is explored using statistical tools.

Statistical Analysis

When the algorithm is executed and clusters are formed, a statistical analysis of the gathered data is done. The clusters and the pre-processed datasets are the foundation of this analysis. The aim in this step is to go beyond a mere data collection and look for the meaning of the patterns found in the clusters.

To properly compare the user trajectories from the user performance and presence self-evaluation, this research focuses on the clusters obtained from the k-means with DTW metric algorithm. This comparison is important to understand not just how users move in the virtual path, but also how this navigation influences their perceived experience and performance. The Chi-squared (χ^2) test is used here as a statistical tool to measure the strength and significance of these relationships. To properly perform this statistical test, data is organized in the code to resemble contingency tables with the cluster labels and the categorical results from the metrics. Multiple tests are required to consider all combinations of cluster labels and metrics. Some comparisons are even done twice, with different level categories, to add confidence to the results.

The Chi Squared test helps reveal whether observed user outcomes across different trajectory clusters occurs by randomness or reflect a significant pattern. If the Chi-squared test yields significant results, it would suggest that the way users navigate through the virtual path is linked to their performance metrics or their perception of the experience. To perform this statistical test, it is essential to formulate a null hypothesis (H_0) which states that there is no association between the clusters and the metrics; this null hypothesis is either rejected or failed to reject, which indicates whether there is an association in the data. The test consists of assuming the null

hypothesis to be true and calculating an expected value for the different metrics in this "no association" scenario. Formula 3.2 shows how the expected value is obtained.

$$\hat{E}_{ij} = \frac{(R_i)(C_j)}{N} \tag{3.2}$$

where

$\hat{E}_{ij}$ is the expected occurrences in the cell in the row i and column j of the contingency table.

R_i is the sum of the values in row i.

C_j is the sum of the values in column j

N is the summation of all values in the contingency table

After calculating the expected value, formula 3.3 shows how this expected value is used to obtain a statistic that measures how much the observed values in each cell of the contingency table deviate from the expected value.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.3}$$

χ^2 represents the chi-squared statistic.

O_i is each observed (actual) value.

E_i is each expected value.

Σ indicates that the operation is done for each cell in the contingency table and then these results are summed.

After obtaining the statistic, another chi-squared value is sought in a chi-squared distribution table. This table [39] consists of rows represented by a value of degrees of freedom (df) and columns represented by significance levels. The df is calculated by multiplying the number of rows in contingency table minus one by the number of columns in contingency table minus one. The appropriate degrees of freedom obtained correspond to the row and the chosen

significance level for this experiment of 0.05 determines the column. The intersection of these is known as the critical value and this critical value is compared to the value obtained with the formula.

If the formula value for χ^2 is greater than or equal to the critical value from the table, the null hypothesis is rejected indicating a possible association between the clusters and the metrics. If the formula χ^2value is less than the critical value, the null hypothesis fails to be rejected, indicating that there is not enough evidence to conclude an association between the clusters and the metrics.

The p-value is obtained to provide a probability-based perspective to the interpretation. This statistic indicates the probability of obtaining the observed results in the null hypothesis scenario. A large p-value indicates that the observed values are likely to occur in a null hypothesis state, supporting it. A p-value smaller than 0.05 suggests that the observed values are unlikely to have occurred by random chance under the null hypothesis, leading to its rejection.

In a consistent, well conducted statistical test, the comparison between chi-squared values and the p-value produces similar results. P-value is used to add more detail to the chi-squared test result.

CHAPTER IV

RESULTS

Database

The final version of the nine databases includes the following:

- Subject ID
- Gender
- Dominant Hand
- Time Steps
- Answers to the 6 survey questions
- Answer to how self-evaluation of difficulty to balance the ball
- Answer to the self-evaluation of "gaming" experience
- Number of times ball was dropped during full path and after a distraction
- X, Y, and Z scaled coordinates for its corresponding trajectory

Gender and Dominant Hand have binary values that represent Male or Female and Right or Left Hand respectively. The time, x,y, and z coordinates are the only variables used to obtain the clusters and these clusters are compared against the other variables (metrics).

Elbow Method

The optimal k-values per situation are shown in Table 4.1 where the three path distractions are shown with their three trajectories.

Table 4.1
K-values obtained from Elbow Method.

Stimuli Event	Trajectory	K-Value
Explosion	Head	7
	Dominant Hand	7
	Non-Dominant Hand	7
Meteor	Head	7
	Dominant Hand	8
	Non-Dominant Hand	5
Birds	Head	7
	Dominant Hand	7
	Non-Dominant Hand	7

The optimal k-value is found within a range from 5 to 8 with the mode being 7. The only event that varies in k-value is Meteor, showing a different value per trajectory.

Trajectory Clusters

The optimal k-value determines the ideal number of clusters to be created in each of the nine scenarios through k-means algorithm with DTW metric. The amount of time series distributed to each cluster is based on their proximity to each other or how similarly the users moved throughout the virtual path. This distribution of trajectories among the optimal amount of clusters is shown in table 4.2.

Table 4.2
The Amount of Time Series (Trajectories) Grouped into the Clusters.

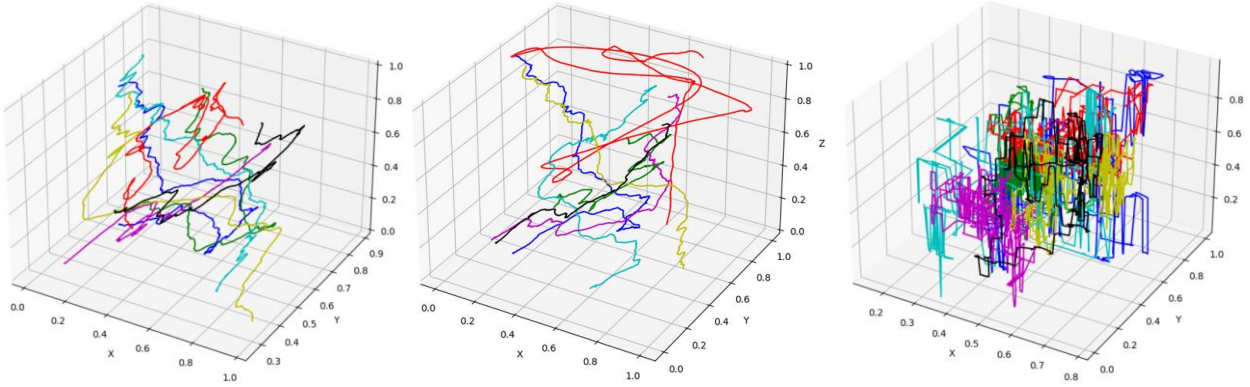
-	-	**Cluster**							
Stimuli Event	**Trajectory**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Explosion	Head	6	7	4	8	13	6	13	-
	Dominant Hand	6	8	1	6	12	9	15	-
	Non-Dominant Hand	3	11	9	4	10	11	9	-
Meteor	Head	10	10	14	9	9	3	2	-
	Dominant Hand	7	5	12	9	3	6	13	2
	Non-Dominant Hand	6	21	4	13	13	-	-	-
Birds	Head	3	10	8	3	12	7	14	-
	Dominant Hand	4	2	6	11	11	8	15	-
	Non-Dominant Hand	8	4	12	6	12	8	7	-

Only one cluster with one trajectory is found. This anomaly occurred in the Explosion-Dominant Hand trajectory. Meteor-Head trajectory seems to have the most balanced distribution except for the last two clusters with fewer time series. Another finding is that even though Meteor-Non-Dominant Hand trajectory has the least (5) clusters, two of them have a smaller number of time series compared to the others.

The clusters generated can be visually represented as centroids of the trajectories or time series. Figures 4.1 through 4.9 illustrate these centroids in three sets, each corresponding to a movement type after a distinct event. To accurately represent the normalized trajectory lengths, the plots are scaled with a three-dimensional spatial distance (x,y, and z coordinates) ranging from 0 to 1.

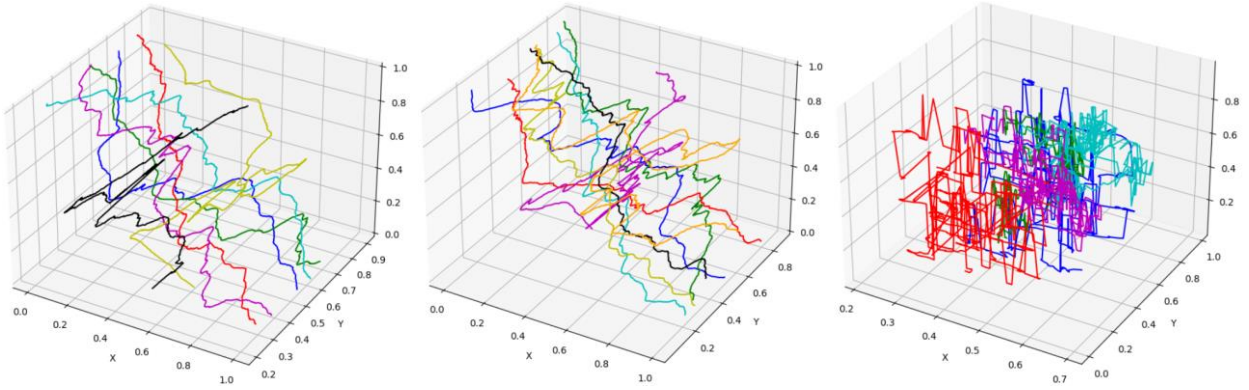
Figures 4.1-4.3
Plotting of normalized trajectory cluster centroids for Explosion event

4.1 Head **4.2 Dominant Hand** **4.3 Non-Dominant Hand**



Figures 4.4-4.6
Plotting of normalized trajectory cluster centroids for Birds event

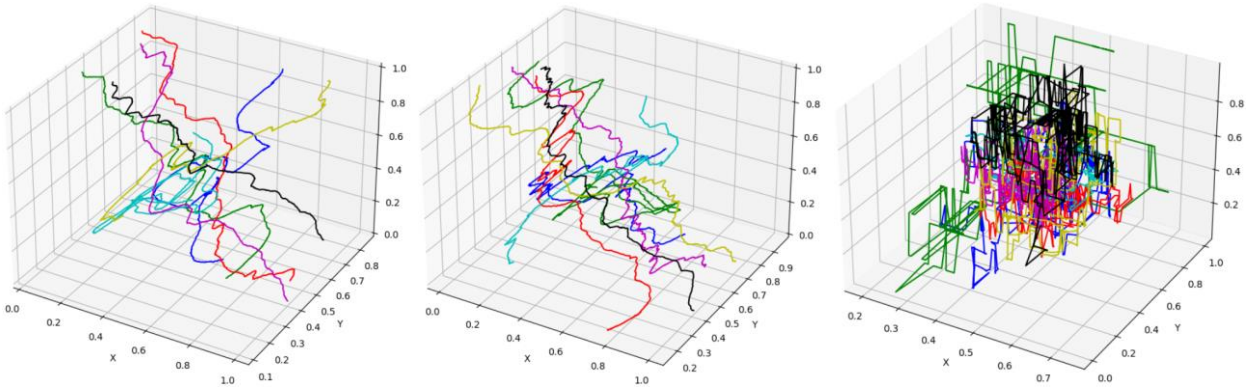
4.4 Head **4.5 Dominant Hand** **4.6 Non-Dominant Hand**



A consistent color mapping is adopted across all nine figures to represent each cluster label and aid in visual comparison. The plots show that head and dominant hand clusters have well-defined clusters with distinct separations among them. In contrast, figures 4.3, 4.6, and 4.9, which show the clusters associated with non-dominant hand trajectories, appear more intermingled.

Figures 4.7-4.9
Plotting of normalized trajectory cluster centroids for Explosion event

4.7 Head **4.8 Dominant Hand** **4.9 Non-Dominant Hand**



Chi-Squared Test Results

Multiple chi-squared tests are conducted to compare the association between the clusters and the different variables or metrics. In each of the nine situations, tests are conducted to assess the association between user gender, trajectory clusters, and user hand dominance. The test comparing user gender and hand dominance yields the same results across the nine situations as these are independent variables; their value does not depend on the trajectories. The dependent variable, clusters, is tested for association between the following:

- The answers to the presence survey - categorized into two separate contingency tables: one with 3 category levels and one with 2 category levels.
- The number of times that users dropped the ball - categorized into two separate contingency tables: one considering the full path and one considering the path section during an individual stimuli event.
- The ‘gamer’ and ‘difficulty to balance ball’ self-perceived scores - organized in 3 category levels each.

Presence Survey: 3-Level

The aggregated scores of the presence survey answers ranging from 1 to 42, are categorized into three levels reflecting presence intensity: low (1-14), neutral (15-28), and high (29-42).

Gender and Hand Dominance Analysis. The association between user gender and survey response is analyzed. The formula χ^2 value is 3.229 and the critical (table) value is 5.991. The p-value is 0.198. For hand dominance analysis, the formula χ^2 value is 0.524 and the critical (table) value is also 5.991 since it shares the df of user gender. The p-value is 0.769.

Clusters Analysis. The number of users whose presence survey responses fell into the three categories are distributed across the different trajectory clusters. Table 4.3 summarizes the statistics obtained.

Table 4.3
Statistics for Clusters-3 Level Presence Association.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	p-Value
Explosion	Head	11.748	21.026	0.466
	Dominant	25.541	21.026	0.0124
	Non-Dominant	6.133	21.026	0.9091
Meteor	Head	6.481	21.026	0.8899
	Dominant	11.117	23.685	0.6767
	Non-Dominant	12.423	15.507	0.1332
Birds	Head	8.46	21.026	0.7482
	Dominant	14.586	21.026	0.2648
	Non-Dominant	13.29	21.026	0.3483

All of the p-values obtained in this association are higher than 0.05, except for the Explosion-Dominant Hand trajectory (0.0124). The chi-squared formula value exceeds the table value for Explosion-Dominant Hand trajectory. The smallest difference between χ^2formula and table value

is found on Meteor-Non-Dominant trajectory; it also has the second smallest p-value (0.1332). The biggest variation between chi-squared values is found in the Explosion-Non-Dominant trajectory.

Presence Survey: 2-Level Categories.

The same presence survey answers are now divided into 2 levels to reflect presence intensity: high (28-42) and low (1-27).

Gender and Hand Dominance Statistics. The formula χ^2 value for user gender is 0.974 and the critical value is 3.841. The p-value is 0.323. The hand dominance statistics are a formula χ^2 value of 0.021, and the critical or table value the same as gender. The p-value is 0.88.

Clusters Statistics. Table 4.4 displays the statistics obtained from conducting the chi-squared test with the presence survey scores divided into 2 categories now.

Table 4.4
Statistics for Clusters-2 Level Presence Association.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	p-Value
Explosion	Head	2.842	12.592	0.828
	Dominant	6.758	12.592	0.343
	Non-Dominant	6.465	12.592	0.373
Meteor	Head	2.549	12.592	0.862
	Dominant	5.082	14.067	0.649
	Non-Dominant	6.459	9.488	0.167
Birds	Head	4.483	12.592	0.611
	Dominant	6.027	12.592	0.42
	Non-Dominant	4.373	12.592	0.626

The biggest deviation between chi-squared values is Meteor-Head trajectory while the smallest deviation, again, is from Meteor-Non-Dominant Hand.

Ball Drops

The amount of times the users dropped the ball during the full path and during an individual event are categorized into 4 levels:

- 1st level is if the ball was dropped only once
- 2nd level is if the ball was dropped 2 times
- 3rd level is if the ball was dropped 3 times
- 4th level represents if the ball was dropped more than 3 times

The amount of times the users dropped ball during a single section (the trajectory after a distraction) did not go over two, therefore only 2 categories are used:

- 1st level if the ball was dropped only once
- 2nd level if the ball was dropped 2 times

Gender and Hand Dominance Statistics. In the full path scenario, which includes the three distractions, the formula chi-squared value for user gender is 8.424, with the table value at 7.815, and a p-value at 0.038. In the single path sections, analyzing the association between the amount of balls dropped and gender, the table value is 3.841, with a chi-squared formula value of 0.482, which is lower than on full path. The corresponding p-value is 0.487, higher than in the full path scenario.

The hand dominance statistics for the Full Path scenario are: the formula χ^2 value of 2.928 and the critical value at 7.815, identical to the gender statistic. The p-value here is 0.402, slightly lower than the 0.497 p-value for Path Section scenario. In Path Section scenario, the χ^2 critical value is shared with gender statistic, and the formula value is 0.461, lower than the value in full path.

Clusters Analysis. Tables 4.5 and 4.6 show the statistics obtained for the χ^2 test for association between the users grouped into the clusters and the amount of ball drops during full path and per section. Table 4.5 displays that the p-value for the Explosion-Dominant Hand is small but not lower than 0.05.

Table 4.5
Statistics for Association between Clusters and Ball Drops during Full Path.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	p-Value
Explosion	Head	11.02	28.869	0.893
	Dominant	28.148	28.869	0.059
	Non-Dominant	18.543	28.869	0.42
Meteor	Head	15.066	28.869	0.657
	Dominant	24.806	32.671	0.255
	Non-Dominant	17.333	21.026	0.137
Birds	Head	12.144	28.869	0.839
	Dominant	16.63	28.869	0.548
	Non-Dominant	13.817	28.869	0.74

In Table 4.6, three scenarios have a p-value below 0.05. Explosion-Non-Dominant Hand's formula value of 21.107 is higher than the table value of 12.592. For the Meteor event, Head and Dominant Hand clusters have the p-values of 0.033 and 0.013 respectively and their formula values are also bigger than the table values.

Table 4.6
Statistics for Association between Clusters and Ball Drops during a Path Section.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	p-Value
Explosion	Head	20.18	12.592	0.063
	Dominant	8.906	12.592	0.71
	Non-Dominant	21.107	12.592	0.048
Meteor	Head	22.372	12.592	0.033
	Dominant	28.197	14.067	0.013
	Non-Dominant	13.014	9.488	0.111
Birds	Head	4.572	12.592	0.599
	Dominant	5.803	12.592	0.445
	Non-Dominant	5.311	12.592	0.504

'Gamer' Level Self-Perception

The self-perceived level of expertise in videogames is divided into three categories:

- Low expertise: 1-2
- Medium: 3-4
- High level of expertise: 5-7

Gender and Hand Dominance Statistics. In comparing these scores with user genders using a chi-squared test, a χ^2 formula value of 17.377, table value of 5.991 and a p-value of 0.00016 are obtained. The association with user hand dominance yields a χ^2formula value of 0.59, with the same table value as gender test, and a p-value of 0.744.

Clusters Statistics. The results of the χ^2 test between cluster distribution and question scores are in Table 4.7.

Table 4.7
Statistics for Clusters and 'Gamer' Self-Perception Association.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	p-Value
Explosion	Head	12.678	21.026	0.39
	Dominant	13.608	21.026	0.326
	Non-Dominant	11.795	21.026	0.462
Meteor	Head	10.401	21.026	0.58
	Dominant	11.736	23.685	0.627
	Non-Dominant	7.311	15.507	0.503
Birds	Head	6.264	21.026	0.902
	Dominant	8.31	21.026	0.744
	Non-Dominant	12.04	21.026	0.442

According to Table 4.7, all situations present p-values above 0.05, the lowest being 0.326 in the Explosion-Dominant clusters. The biggest deviation between observed values and expected values is the trajectory of Birds-Dominant Hand.

'Balancing Ball Difficulty' Self-Perception

The users were asked at the end of the experiment, how would they rate the difficulty balancing the ball while walking in the virtual path and the answers were categorized the same as the 'Gamer' score: low expertise (1-2), medium (3-4), and high level of expertise (5-7).

Gender and Hand Dominance Statistics. In the gender metric association, the chi-squared formula value is 4.232, with a table value of 5.991, and a p-value of 0.12. When comparing 'Balancing Ball Difficulty' to amount of users with different hand dominance, the χ^2 formula value is 3.33, with the same table value of 5.991, and a p-value of 0.188.

Clusters Statistics. The χ^2values obtained when analyzing the association between users' perceived the difficulty of balancing the ball and the clusters of their trajectory are shown in Table 4.8.

Table 4.8 shows that Explosion-Dominant trajectory has the only low p-value (0.013). The largest deviation between observed values and expected values is the trajectory of Birds-Dominant Hand.

Table 4.8
Statistics for Clusters and 'Difficulty to Balance Ball' Self-Perception Association.

Stimuli	Trajectory	χ^2 Formula Value	χ^2 Table Value	P-Value
Explosion	Head	11.209	21.026	0.511
	Dominant	11.482	21.026	0.538
	Non-Dominant	8.468	21.026	0.747
Meteor	Head	10.544	21.026	0.568
	Dominant	28.129	23.685	0.013
	Non-Dominant	9.712	15.507	0.285
Birds	Head	13.642	21.026	0.324
	Dominant	16.072	21.026	0.187
	Non-Dominant	19.424	21.026	0.078

CHAPTER V

DISCUSSIONS

Database Complexity

The goal of the study was to accurate cluster spatiotemporal trajectories to statistically test them with independent variables, and this analysis required the modeling of the dataset. The raw state of the initial datasets was not ideal for analysis due to several reasons:

- Users that walked the virtual path were identified differently across the available data.
- It contained data that would not be part of the experiment.
- The spatiotemporal coordinates were in different ranges.
- The independent variables were located in a separate dataset.
- The spatiotemporal trajectory coordinates were scattered along several worksheets.
- The raw coordinates cover the complete length of Gracia's [1] experiment, including time ranges with useless data.

Another aspect when restructuring the dataset is to consider how the algorithm interprets this data. When OPTICS and DBSCAN algorithms were considered and tried, there was no focus on the order of the temporal coordinates in the dataset since both have no input that tells them to look for trajectory length. OPTICS looks for a minimum density value

The K-means algorithm with DTW metric measures the distance between time series in a defined space. By rescaling the spatial coordinates, we prevented the algorithm from giving

more weight to data points with disproportionately large values, ensuring each coordinate contributed equally to the calculation and improving the quality of the clusters. By transforming the tme measurement to time steps, the complexity of the temporal data was reduced along with the computational time to interpret it.

Clusters

Time Series Distribution Insights

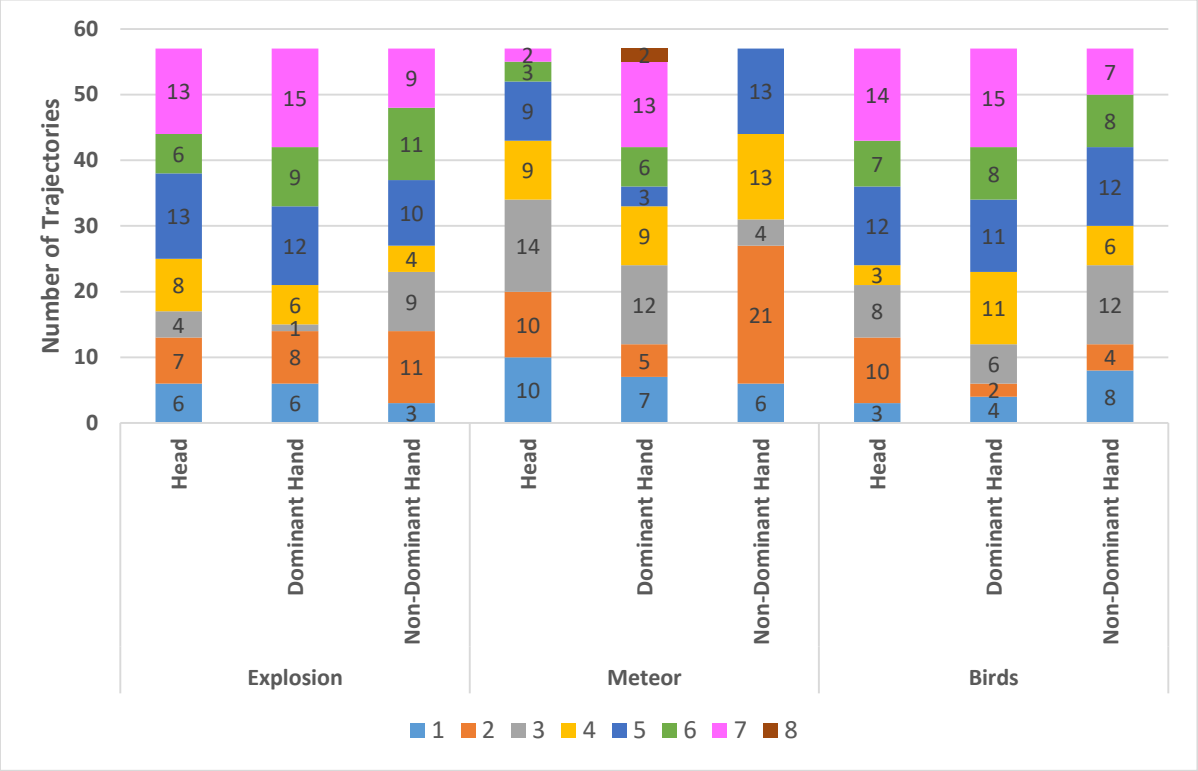
The clusters created capture the full user trajectories, increasing the reliability of the analysis. Since clusters are within the same spatiotemporal range, their visual trajectory representations are easy to distinguish. These clear trajectories make it simpler to understand how users moved in each situation. Table 4.2 provides the amount of trajectories distributed among the clusters in each scenario. Additionally, Figure 5.1, visually breaks down this distribution: this visual representation helps in quickly identifying trends or anomalies in the trajectory data.

The first observation to consider from Figure 5.1 is that the number of clusters among the situations is similar, ranging from 5 to 8; this can be caused by the similar data structure and the same amount of data points that each situation has.

Following Explosion distraction, the head movement data peaks in the 5^{th} and 7^{th} clusters. In contrast to the head trajectories in the other two events, ‘Explosion’ displays drastic concentration peaks. Another observation from the distribution is found on the 9^{th} bar. This bar that represents non-dominant hand trajectory after ‘Birds’ event, shows the most balanced clusters since they look more evenly spaced.

Figure 5.1

Time Series Distribution Across Clusters



A significant finding from Figure 5.1 is the observed similarity in non-dominant hand movements among users following the Meteor event; cluster 2 has the highest concentration of time series. Explosion and Birds dominant hand movements have the second highest concentrations of trajectories with 15 in cluster 7.

A final observation found in Figure 5.1 is that the head movement after the 'Meteor' event was high. Five out of the seven clusters show a notable count for head trajectories.

A couple of outliers are present in the distribution. These are clusters with very few time series. For example, the dominant hand 3rd cluster after 'Explosion' event has only one trajectory. Such time series can be excluded in a future analysis to execute the algorithms again and explore differences in the distribution.

Standard Deviation Insights

The analysis of the standard deviation (σ) in Figure 5.1 offers a more complete view of the distribution patterns. Table 5.1 lists the standard deviation values for each scenario, these are further explained in the following sections.

Head Trajectory. The head movements, or trajectories, after the Meteor and Birds events have an identical high variability while the Explosion event has a modre moderate variability.

Dominant Hand Trajectory. There is moderately high variability in the Explosion and Birds categories. Their standard deviation indicates distinct ranges of movements, but not as varied as the scenarios with higher standard deviation value.

Non-Dominant Hand Trajectory. The standard deviation for the distribution of time series among clusters after the Explosion and Birds events shows a low to moderate variability. The users' non-dominant hand trajectories showed similar behaviors. The variability after the Meteor event is very high, indicating a wide range of trajectory patterns.

Table 5.1
Standard Deviation for the Time Series Distribution at each Situation.

Stimuli Event	Trajectory	Standard Deviation
Explosion	Head	3.53
	Dominant Hand	4.52
	Non-Dominant Hand	3.29
Meteor	Head	4.22
	Dominant Hand	3.67
	Non-Dominant Hand	7.82
Birds	Head	4.22
	Dominant Hand	4.53
	Non-Dominant Hand	2.97

Chi-Squared Test Interpretations

The results shown in Chapter IV for the Chi-Squared Tests provide a measure of association between the data. A null hypothesis that indicated there was no association between data was assumed and the statistical test would reject or fail to reject this assumption.

Presence Survey

In almost all chi-squared tests evaluating the association between the Presence survey scores and the independent variables or clusters, the table value obtained was higher than the value obtained by formula. This indicates that there is not enough evidence to support an association between the data, therefore, the null hypothesis fails to be rejected.

Only one test for association resulted in a p-value lower than 0.05. The chi-squared test for Dominant hand clusters after the Explosion event and presence scores resulted in a p-value of 0.0124 which allows for the rejection of the null hypothesis in this association. The remaining tests showed a high p-value. This high p-value complements the chi-squared test results by suggesting that the observed values are likely to occur in a null hypothesis scenario.

Ball Drops

The gender association test with amount of ball drops during the full path had a chi-squared formula value higher than the table value. The null hypothesis is, therefore, rejected, showing association between data. The p-value of 0.038 suggests that the observed values are unlikely to have occurred by chance, supporting the chi-squared test result.

The ball drops per path section showed an association with body trajectories. Dominant hand and head trajectories in the Meteor event have an association with the amount of times the ball was dropped per path section based on the chi-squared test results in Table 4.6. This did not affect overall stability since there is no relationship with non-dominant hand movement. The

Explosion event also had a chi-squared result that rejected null hypothesis for association between head and dominant hand with ball drops per event.

An interesting observation is the lack of association between ball drops and trajectories in the Birds event. In contrast to the other two events, Birds did not seem to cause an effect on ball drops.

Gamer Level

The chi-squared test result and significantly low p-value of 0.000016 obtained from the association between gender and user self-perceived level of "gamer" suggest a statistical relationship between them.

The χ^2 formula value in Table 4.7 for the association between "gamer" auto-declared level and the formed clusters is higher than the χ^2 table value. This difference means that there is not enough evidence to support the null hypothesis, rejecting it. The low p-value of 0.0164 also suggest that the observed values are not likely to be obtained in a situation with no data association.

Difficulty to Balance Ball

In this category, only one situation showed a chi-squared result that caused the null hypothesis to be rejected due to insufficient evidence to support. After the Explosion event, there is association between the dominant hand movements and how difficult users perceived it was to balance the ball and the p-value of 0.0136 supports it.

Future Research

Now that the clusters are available, performing more statistical tests on them can help provide more findings and interpretations.

Multivariate Analysis of Variance (MANOVA) can overview the impact of the independent variables simultaneously on the count of trajectories in each cluster. While chi-squared tests were done individually per situation, MANOVA can handle all situations at the same time and provide insights to how they interact and influence output collectively. This can provide a deeper understanding of how the stimuli events and users' demographics affect the trajectories in the virtual environment.

Another potential analysis that can be performed is regression analysis. Regression analysis can help understand how strongly each metric can predict user trajectories or clusters.

Summary

The spatiotemporal user trajectory clusters within in a virtual path were analyzed using three different methods. Pattern seeking, standard deviation, and chi-quared statistical test complement each other to produce reliable results. One of the findings is that after the Explosion stimuli, participants' dominant and non-dominant hand movements varied widely. This observation is supported by high standard deviation values and the chi-squared test which suggested that these movements were important factors in the amount of ball drops

The goal of this research was achieved. Clusters were formed using the spatiotemporal trajectories taken by users on a virtual environment and these clusters gave meaningful insights into user performance and user's demographics.

REFERENCES

[1] Gracia de Luna, D., Roel, T., Butler, A., Tomai, E., Timmer, D., & Dumitru, C. (2019). *A STUDY OF HUMAN BALANCE AND COORDINATION USING A HEAD MOUNTED DISPLAY*.

[2] Gracia de Luna, D., Tijernia, R., Butler, A., Tomai, E., Timmer, D., & Caruntu, D. (2020, August). A STUDY OF HUMAN BALANCE AND COORDINATION USING A HEAD MOUNTED DISPLAY. *ASME 2020 International Design Engineering Technical Conference and Computers and Information in Engineering*.

[3] Gracia de Luna, D., Butler, A., & Timmer, D. (2021). A Study of Human Balance and Coordination Using a Head Mounted Display. *Journal of Computing and Information Science in Engineering, 21*.

[4] Hament, B., Cater, A., & Y. Oh, P. (2017). Coupling Virtual Reality and Motion Platforms for Snowboard Training. *14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI).*

[5] Pengnate, S., Riggins, F. J., & Zhang, L. (2020). Understanding Users' Engagement and Responses in 3D Virtual Reality: The Influence of Presence on User Value. *Interacting with Computers*, *32*(2), 103–117. https://doi.org/10.1093/iwc/iwaa008

[6] Slater, M., Usoh, M., & Steed, A. (1994). Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, *3*(2), 130–144. https://doi.org/10.1162/pres.1994.3.2.130

[7] Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using Presence Questionnaires in Reality. *Presence*, *9*(5), 497–503.

[8] Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis*. Thomson.

[9] Zhu, Y., Wu, Y., Min, G., Zomaya, A., & Hu, F. (2018). *A Survey of Big Data and Computational Intelligence in Networking*.

[10] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

[11] Ozer, M., & Cebeci, U. (2019). Affective design using big data within the context of online shopping. *Journal of Engineering Design*, *30*(8–9), 368–384. https://doi.org/10.1080/09544828.2019.1656803

[12] Bulagang, A. F., Mountstephens, J., & Teo, J. (2021). Multiclass emotion prediction using heart rate and virtual reality stimuli. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-020-00401-x

[13] Hasenbein, L., Stark, P., Trautwein, U., Gao, H., Kasneci, E., & Göllner, R. (2023). Investigating social comparison behaviour in an immersive virtual reality classroom based on eye-movement data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-41704-2

[14] Birenboim, A., Dijst, M., Ettema, D., de Kruijf, J., de Leeuw, G., & Dogterom, N. (2019). The utilization of immersive virtual environments for the investigation of environmental preferences. *Landscape and Urban Planning*, *189*, 129–138. https://doi.org/10.1016/j.landurbplan.2019.04.011

[15] Bouchard, S., & St-Jacques, J. (2008). Anxiety Increases the Feeling of Presence in Virtual Reality by Bouchard. *Presence*, *17*(4), 376–391.

[16] Chuan, A., Qian, J., Bogdanovych, A., Kumar, A., McKendrick, M., & McLeod, G. (2023). Design and validation of a virtual reality trainer for ultrasound-guided regional anaesthesia. *Anaesthesia*, *78*(6), 739–746. https://doi.org/10.1111/anae.16015

[17] Xia, D., Jiang, S., Yang, N., Hu, Y., Li, Y., Li, H., & Wang, L. (2021). Discovering spatiotemporal characteristics of passenger travel with mobile trajectory big data. *Physica A: Statistical Mechanics and Its Applications*, *578*. https://doi.org/10.1016/j.physa.2021.126056

[18] Ansari, M. Y., Ahmad, A., Khan, S. S., Bhushan, G., & Mainuddin. (2020). Spatiotemporal clustering: a review. *Artificial Intelligence Review*, *53*(4), 2381–2423. https://doi.org/10.1007/s10462-019-09736-1

[19] Peng, Y., Deng, M., Yu, Y., Hu, Z., Wang, K., Wang, X., Yi, S., & Deng, G. (2023). *Analysis of moose motion trajectory after bullet train-moose collisions*. https://doi.org/10.1016/j.eng□ailanal.2023.107373

[20] Zhao, J., & Li, Z. (2022). Recognition of Volleyball Player's Arm Motion Trajectory and Muscle Injury Mechanism Analysis Based upon Neural Network Model. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/8114740

[21] Song, H. S., Lu, S. N., Ma, X., Yang, Y., Liu, X. Q., & Zhang, P. (2014). Vehicle behavior analysis using target motion trajectories. *IEEE Transactions on Vehicular Technology*, *63*(8), 3580–3591. https://doi.org/10.1109/TVT.2014.2307958

[22] Nanni, M., & Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, *27*(3), 267–289. https://doi.org/10.1007/s10844-006-9953-7

[23] Yuan, G., Sun, P., Zhao, J., Li, D., & Wang, C. (2017). A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, *47*(1), 123–144. https://doi.org/10.1007/s10462-016-9477-7

[24] Han, Kamber, M., & Pei, J. (2012). Data mining concepts and techniques (3rd ed.). Elsevier.

[25] Li, H., Liu, J., Liu, R. W., Xiong, N., Wu, K., & Kim, T. H. (2017). A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis. *Sensors (Switzerland)*, *17*(8). https://doi.org/10.3390/s17081792

[26] Agrawal, K. P., Garg, S., Sharma, S., & Patel, P. (2016). Development and validation of OPTICS based spatio-temporal clustering technique. *Information Sciences*, *369*, 388–401. https://doi.org/10.1016/j.ins.2016.06.048

[27] Bushra, A. A., & Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access*, *9*, 87918–87935. https://doi.org/10.1109/ACCESS.2021.3089036

[28] Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

[29] Chimwayi, K. B., & Anuradha, J. (2018). Clustering West Nile Virus Spatio-temporal data using ST-DBSCAN. *Procedia Computer Science*, *132*, 1218–1227. https://doi.org/10.1016/j.procs.2018.05.037

[30] Malhan, A. (2017). *ST-OPTICS: A spatial-temporal clustering algorithm with time recommendations for taxi services*.

[31] Zhuang, S., & Chen, C. (2021, July). Aerial battlefield targets grouping based on DTW-DBSCAN algorithm-1. *40th Chinese Control Conference*.

[32] Chen, T., Shi, X., & Wong, Y. D. (2021). A lane-changing risk profile analysis method based on time-series clustering. *Physica A: Statistical Mechanics and Its Applications*, *565*. https://doi.org/10.1016/j.physa.2020.125567

[33] Zhao, L., & Shi, G. (2019). A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. *Ocean Engineering*, *172*, 456–467. https://doi.org/10.1016/j.oceaneng.2018.12.019

[34] Dupas, R., Tavenard, R., Fovet, O., Gilliet, N., Grimaldi, C., & Gascuel-Odoux, C. (2015). Identifying seasonal patterns of phosphorus storm dynamics with dynamic time warping. *Water Resources Research*, *51*(11), 8868–8882. https://doi.org/10.1002/2015WR017338

[35] Xing, L., & Sela, L. (2019). Unsteady pressure patterns discovery from high-frequency sensing in water distribution systems. *Water Research*, *158*, 291–300. https://doi.org/10.1016/j.watres.2019.03.051

[36] Zhang, E., Masoud, N., Bandegi, M., & Malhan, R. K. (2022). Predicting Risky Driving in a Connected Vehicle Environment. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 17177–17188. https://doi.org/10.1109/TITS.2022.3170859

[37] Chen, T., Shi, X., & Wong, Y. D. (2021). A lane-changing risk profile analysis method based on time-series clustering. *Physica A: Statistical Mechanics and Its Applications*, *565*. https://doi.org/10.1016/j.physa.2020.125567

[38] Jang, M., Han, M.-S., Kim, J.-H., & Yang, H.-S. (2011). *Dynamic Time Warping-Based K-Means Clustering for Accelerometer-Based Handwriting Recognition*

[39] James, R. K. (n.d.). Chi-squared table. Richland Community College. Retrieved from https://people.richland.edu/james/lecture/m170/tbl-chi.html

APPENDIX

APPENDIX

WORKSHEET MERGE VBA

```
Sub CombineWorksheets()
    Dim J As Integer
    Dim DestRow As Long
    Dim ws As Worksheet
    ' Create a new worksheet to store the combined data
    Set ws = ThisWorkbook.Worksheets.Add(Before:=ThisWorkbook.Sheets(1))
    ws.Name = "Combined"
    ' Copy the headers from the second sheet to the combined sheet
    ThisWorkbook.Sheets(2).Rows(1).Copy Destination:=ws.Rows(1)
    ' Initialize the destination row (headers is 1)
    DestRow = 2
    ' Loop through each sheet and copy all its data and paste on the combined
sheet
    For J = 2 To ThisWorkbook.Sheets.Count
        With ThisWorkbook.Sheets(J)
            ' Find the last row with data in the current sheet
            LastRow = .Cells(.Rows.Count, "A").End(xlUp).Row
            ' Copy data excluding headers
            If LastRow > 1 Then
                .Range("A2:Z" & LastRow).Copy Destination:=ws.Cells(DestRow,
1)
                DestRow = ws.Cells(ws.Rows.Count, "A").End(xlUp).Row + 1
            End If
        End With
    Next J
End Sub
```

APPENDIX

ELBOW METHOD CODE

```
import numpy as np
import pandas as pd
from tslearn.clustering import TimeSeriesKMeans
import matplotlib.pyplot as plt

# Load the data
df = pd.read_excel('path_to_your_excel_file.xlsx')  # Replace with your file path

num_timeseries = 57
time_series_data = []

# Extract each time series and add to list
for i in range(num_timeseries):
    # Extract the MTS for each time series
    X = df.iloc[i*720 : (i+1)*720, 16:19].values  # Adjust columns as needed
    time_series_data.append(X)

time_series_data = np.squeeze(time_series_data)

# Define the range of clusters to test
clusters_range = range(2, 11)  # clusters from 2 to 10
sse = []  # list to hold the sum of squared errors for each cluster size

# Loop over the clusters
for k in clusters_range:
    # Initialize a time series k-means model
    km = TimeSeriesKMeans(n_clusters=k, metric="dtw", random_state=1)
```

```
    # Fit the model to the data
    km.fit(time_series_data)

    # Append the inertia (SSE) to the list
    sse.append(km.inertia_)

# Plot SSE against k
plt.figure(figsize=(6, 6))
plt.plot(clusters_range, sse, '-o')
plt.xlabel('Number of clusters *k*')
plt.ylabel('Sum of squared distance')
plt.title('Elbow Method For Optimal k')
plt.show()
```

APPENDIX

K-MEANS DTW CODE

```
# Presence 3-Level Gender


import scipy.stats as stats

# histogram of presence

plt.hist(df2['Q-Total'])

plt.show()

#

df2['Presence'] = pd.cut(x=df2['Q-Total'], bins=[1,14,28,42],right=False)

# Contingency Table

comparison1 = pd.crosstab(df2['Presence'],df2['Gender'])

print(comparison1)

#

# are these variables independent?

chi2, p, dof, ex = stats.chi2_contingency(comparison1)


print(f'Chi_square value {chi2}\n\np value {p}\n\ndegrees of freedom
{dof}\n\n expected {ex}')
```

BIOGRAPHICAL SKETCH

Martín Alejandro Galicia Avila was born in Matamoros, Tamaulipas in 1005. Martín earned his Bachelor's of.Science. in Manufacturing Engineering in the University of Texas Rio Grande Valley (UTRGV) in 2020. During his bachelor's studies, he was involved in several student organizations such as: Society for Manufacturing Engineers, Aero-Design, Society of Hispanic Engineer Professionals, and Rocket Launchers. He was also invited to structure the base of Society for Women Engineers due to his student organization experience. He graduated in 2023 with a Master's of Science in Engineering Management in UTRGV as well. As he completed his bachelor's and master's, Martín completed five full time internships. Delphi Automotive, Staples Inc, Panasonic Automotive, Reyes Automotive, and a full time in Tesla are his experiences. His email is martin_g4@yahoo.com.