

American University in Cairo

AUC Knowledge Fountain

Theses and Dissertations

Student Research

Winter 1-31-2022

A Framework for Real-time Spatial Labor Data Analytics from Construction Sites

Hoda Abouorban
hoda94@aucegypt.edu

Follow this and additional works at: <https://fount.aucegypt.edu/etds>

Recommended Citation

APA Citation

Abouorban, H. (2022). *A Framework for Real-time Spatial Labor Data Analytics from Construction Sites* [Master's Thesis, the American University in Cairo]. AUC Knowledge Fountain.
<https://fount.aucegypt.edu/etds/1883>

MLA Citation

Abouorban, Hoda. *A Framework for Real-time Spatial Labor Data Analytics from Construction Sites*. 2022. American University in Cairo, Master's Thesis. *AUC Knowledge Fountain*.
<https://fount.aucegypt.edu/etds/1883>

This Master's Thesis is brought to you for free and open access by the Student Research at AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact thesisadmin@aucegypt.edu.



THE AMERICAN UNIVERSITY IN CAIRO

الجامعة الأمريكية بالقاهرة

SCHOOL OF SCIENCES AND ENGINEERING

A FRAMEWORK FOR REAL-TIME SPATIAL LABOR DATA ANALYTICS FROM CONSTRUCTION SITES

A Thesis Submitted to

The Department of Construction Engineering

In partial fulfillment of the requirements for the degree of

Master of Science in Construction Management

By:

Hoda Khaled Abouorban

Under the Supervision of:

Dr. Khaled Nassar

Professor

Department of Construction

Engineering

The American University in Cairo

Dr. Elkhayam Dorra

Adjunct Assistant Professor

Department of Construction

Engineering

The American University in Cairo

December, 2021

TABLE OF CONTENTS

1. CHAPTER 1 – INTRODUCTION	17
1.1 Problem Statement	17
1.2 Lean Approach in Construction	17
1.3 Real-time Monitoring on Construction Sites	19
1.4 Research Objectives	20
1.5 Scope of Work.....	20
1.6 Research Methodology.....	21
1.7 Thesis Organization	22
2. CHAPTER 2 – LITERATURE REVIEW	24
2.1 Waste Allocation (Lean Construction).....	24
2.2 Real-time Monitoring Systems - (RTMSs).....	28
2.2.1 <i>RTMS Application for Health & Safety</i>	29
2.2.2 <i>RTMS Application for Workers' and Equipment's Productivity</i>	35
2.3 Research Gap	47
3. CHAPTER 3 – FRAMEWORK DEVELOPMENT FOR FIRST ORDER ANALYSIS	49
3.1 Stage 1 – Data Collection.....	49
3.1.1 <i>Site Geographic Data Collection – (SGD)</i>	50
3.1.2 <i>Site Periodic Data Collection – (SPD)</i>	52
3.1.3 <i>Real-time Geospatial Data Collection – (RGSD)</i>	54
3.2 Stage 2 – Data Preparation and Cleaning.....	57
3.2.1 <i>Site Geographical Data Preparation - (SGDP)</i>	57
3.2.2 <i>Site Periodic Data Preparation - (SPDP)</i>	59
3.2.3 <i>Real-time Geospatial Data Preparation - (RGSDP)</i>	60
3.3 Stage 3 – First Order Data Analysis (FODA)	62
3.3.1 <i>Analysis Techniques</i>	62
3.3.2 <i>Analysis Techniques Implementation and Outputs</i>	80

4. CHAPTER 4 – FRAMEWORK IMPLEMENTATION FOR SECOND ORDER ANALYSIS.....	98
4.1 Stage 1 - Data Generation and Visualization	99
4.1.1 Site Geographic Data – (SGD).....	100
4.1.2 Site Periodic Data – (SPD).....	102
4.1.3 Real-time Geospatial Data – (RGSD)	104
4.2 Stage 2 – First Order Data Analysis (FODA)	107
4.3 Stage 3 – Second Order Data Analysis (SOCDA).....	125
4.3.1 Correlation between Safety Performance and Productivity.....	125
4.3.2 Correlation between Quality Performance and Productivity.....	130
4.3.3 Other Site Parameters and Productivity.....	135
4.4 Summary of Findings.....	138
5. CHAPTER 5 – CASE STUDY.....	139
5.1 Project Information	139
5.2 Framework Implementation.....	139
5.2.1 Data Collection.....	141
5.2.2 Data Preparation and Cleaning.....	143
5.2.3 First Order Data Analysis	147
6. CHAPTER 6 – CONCLUSION & RECOMMENDATIONS.....	158
6.1 Research Summary and Overview	158
6.2 Research Contributions	159
6.3 Recommendations for Future Research	159
7. REFERENCES.....	161
8. APPENDIX A – PYTHON ALGORITHM.....	163

LIST OF FIGURES

Figure 1-1: Value-stream Maps for Concrete Formwork Activity (KO & KUO, 2015)	18
Figure 1-2: Process Map for Concrete Formwork Activity (KO & KUO, 2015).....	18
Figure 2-1: Waste Generated due to the Nature of Operations (Nikakhtar et. al., 2015)	25
Figure 2-2: Process Waste Comparison between the Real-world and Lean Model (Nikakhtar et. al., 2015)	26
Figure 2-3: Resources Assigned to Activities in the Reinforcement Operation (Nikakhtar et. al., 2015)	26
Figure 2-4: Rebar Activity Process Map (Nikakhtar et. al., 2015)	27
Figure 2-5: VSM Output (Wang et al., 2015)	28
Figure 2-6: Layout of an IoT-based wireless monitoring platform for deep large underground caverns (Zhang et. al., 2021)	32
Figure 2-7: Remote Wireless Transmission Model (Zhang et. al., 2021).....	33
Figure 2-8: Architecture of the real-time and online analysis and early-warning safety system (Zhang et. al., 2021).....	34
Figure 2-9: Architecture of the IoT-based Wireless Monitoring System in Underground Caverns (Zhang et. al., 2021).....	35
Figure 2-10: Conceptual Framework for Integrating Time-lapsing or Videoing (Teizer, 2015)	37
Figure 2-11: Data Aspects to Consider in Knowledge based Decision Making (Teizer, 2015)	38
Figure 2-12: Output of Visual Tracking (Teizer, 2015)	38
Figure 2-13: Real-time Monitoring System Framework (Jiang et al., 2015).....	39
Figure 2-14: ZTE Phone Application Interface (Jiang et al., 2015)	40
Figure 2-15: Web-based Management Application Output Interface (Jiang et al., 2015)	41
Figure 2-16: Sensing Technologies (Calvetti et al.,2020)	42
Figure 2-17: Data Gathered by Sensing Technologies (Calvetti et al.,2020)	43
Figure 2-18: Motion Productivity Specifications and Patterns (Calvetti et al.,2020)..	44
Figure 2-19: Flowchart to Increase Efficiency (Calvetti et al.,2020)	44
Figure 2-20: Productivity Modelling Chart (Calvetti et al.,2020)	46

Figure 3-1: Research Framework Stages	49
Figure 3-2: Google Earth Interface	50
Figure 3-3: Typical Construction Site Layout Plan (The Constructor)	51
Figure 3-4: Exported GPS Coordinates from Google Earth	52
Figure 3-5: Smart Phone GPS Tracking Application (iOS Apple Store)	55
Figure 3-6: Output of using Smart Phone GPS Tracking Application (myTracks iOS Application Interface)	56
Figure 3-7: gpx File Sample	56
Figure 3-8: Area Unique Code Development	58
Figure 3-9: Plot of Moving Average.....	64
Figure 3-10: Mean and Median Centers of Fire Stations in a City (Yuan et. al., 2020)	66
Figure 3-11:Standard Deviation Ellipse of Fire Stations in a City (Yuan et. al., 2020)	67
Figure 3-12: Poisson Distribution for Complete Spatial Randomness (Yuan et. al., 2020)	68
Figure 3-13: G-function Plot (Yuan et. al., 2020).....	69
Figure 3-14: BIRCH Clustering CF-Tree Construction.....	71
Figure 3-15: Birch Clustering Sample Output (Thecleverprogrammer).....	72
Figure 3-16: Quadrat Density of Fire Stations in a City (Yuan et. al., 2020).....	72
Figure 3-17: Normal Distribution Graph	74
Figure 3-18: KDE Bump Smoothing (Chen, 2017)	74
Figure 3-19: Bandwidth Effect on KDE (Chen, 2017)	75
Figure 3-20: KDE Final Plot (Waskom, 2021).....	76
Figure 3-21: KDE Bivariate Contour Plot (Waskom, 2021)	76
Figure 3-22: Heatmap of Fire Stations in a City (Yuan et. al., 2020).....	78
Figure 3-23: Voronoi Diagramme (Elm Packages)	79
Figure 3-24: Heatmap Sample (Waskom, 2021)	82
Figure 3-25: Heatmap Sample (Waskom, 2021)	85
Figure 3-26: Flowchart for Workers' Time Distribution Calculation	95
Figure 3-27: Detailed First Order Analysis Framework Composition	97
Figure 4-1: Construction Site Layout for Random SGD	102
Figure 4-2: Scatterplot of Workers' Coordinates	105

Figure 4-3: Scatterplot of Aggregated Workers' Coordinates	106
Figure 4-4: Line plot of Aggregated Workers' Daily Tracks.....	106
Figure 4-5: Mean, Median, and Standard Deviation of Safety Incidents on Site	108
Figure 4-6: Moving Average Regression of Safety Incidents on Site	109
Figure 4-7: Probability Density Plot of Safety Incidents on Site.....	110
Figure 4-8: Safety Risk Zones - Heatmap of Safety Incidents on Site	110
Figure 4-9: Mean, Median, and Standard Deviation of Inspection Requests on Site	111
Figure 4-10: Moving Average Linear Regression of Inspection Requests on Site....	112
Figure 4-11: Probability Density Plot of Inspection Requests on Site	113
Figure 4-12: Quality Zones - Heatmap of Inspection Requests on Site	114
Figure 4-13: Mean Center of Workers on Site.....	115
Figure 4-14: Mean Center and Median of Workers on Site.....	116
Figure 4-15: Standard Distance Circle of Workers on Site	116
Figure 4-16: Standard Deviation Ellipse of Workers on Site	117
Figure 4-17: G-function Plot for Workers on Site	118
Figure 4-18: G-function Envelope Plot for Workers on Site.....	119
Figure 4-19: BIRCH Clustering of Workers on Site	120
Figure 4-20: Quadrat Density of Workers on Site	121
Figure 4-21: Voronoi Diagram of Workers on Site	122
Figure 4-22: Productivity Zones - Heatmap of Workers' Density on Site	123
Figure 4-23: Workers' Time Distribution on Site	124
Figure 4-24: Workers' Density vs Site Safety Zones	126
Figure 4-25: Workers' Central Tendencies vs Site Safety Zones	127
Figure 4-26: Working Activity vs Site Safety Behavior.....	129
Figure 4-27: Workers' Density vs Site Quality Zones	131
Figure 4-28: Workers' Central Tendencies vs Site Quality Zones.....	132
Figure 4-29: Working Activity vs Site Quality Behavior.....	134
Figure 5-1: Framework Implementation Flowchart.....	140
Figure 5-2: View of Construction Site obtained from Google Earth.....	141
Figure 5-3: Extracted Bounding Coordinates of Site Areas	141
Figure 5-4: gpx File Extracted for a Single Engineer on Site.....	142
Figure 5-5: The Engineer's Track on Site Viewed Using Google Earth.....	143
Figure 5-6: Engineers' Coordinates on Site over the 10-day Working Period.....	147

Figure 5-7: Engineers' Tracks on Site over the 10-day Working Period	148
Figure 5-8: Engineers' Coordinates after RDP Algorithm Application	149
Figure 5-9: Spatial Mean Centers of Engineers	150
Figure 5-10: Median Centers of Engineers' Coordinates	151
Figure 5-11: Standard Circles for Engineers' Coordinates	151
Figure 5-12: Standard Deviational Ellipse of Engineers' Coordinates	152
Figure 5-13: G-function Plots for Engineers' Coordinates	153
Figure 5-14: G-function Envelope Plots for Engineers' Coordinates	153
Figure 5-15: Clustering of Engineers' Coordinates.....	154
Figure 5-16: Quadrat Density of Engineers' Coordinates	155
Figure 5-17: Voronoi Diagrams of Engineers' Coordinates	155
Figure 5-18: Productivity Zones of Engineers on Site.....	156
Figure 5-19: Engineers' Time Distribution	157

LIST OF TABLES

Table 3-1: Safety Incident Log	53
Table 3-2: Inspection Requests Log	54
Table 3-3: Site Geographic Data Tabulation Format.....	57
Table 3-4: Site Geographic Data Sample.....	59
Table 3-5: Grouped Safety Data Tabulation.....	59
Table 3-6: Grouped Inspection Data Tabulation	60
Table 3-7: Tabulation of Real-time Data After gpx Splitting.....	60
Table 3-8: Sample of Real-time Data Tabulated	62
Table 3-9: Analysis Implementation Matrix.....	79
Table 3-10: Output of Workers' Time Distribution	96
Table 4-1: Second Order Analysis Matrix	99
Table 4-2: Randomly Generated SGD.....	100
Table 4-3: Dimensions of Site Areas in meters	101
Table 4-4: Randomly Generated SPD - Safety Records.....	103
Table 4-5: Randomly Generated SPD - Inspection Requests	103
Table 4-6: Sample of Randomly Generated RGSD.....	104
Table 5-1: Transformed Bounding Coordinates of Site Areas	143
Table 5-2: Sample of Engineers' Coordinates.....	146

LIST OF ABBREVIATIONS

SYMBOL	DESCRIPTION
BBS	Behavior-Based Safety
BIM	Building Information Models
BIRCH	Balanced Iterative Reducing and Clustering
CF	Clustering Feature
CP	Concrete Pouring
CSLP	Construction Site Layout Plan
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ED	Emergency Department
EET	Electric/Electronic Tools
ESRI	Environmental Systems Research Institute
EW	Electrical Works
FHP	Free Hand Performing
FODA	First Order Data Analysis
FR	Foreman
FW	Formwork
GIS	Geographic Information System
GPS	Global Positioning Systems
HAZD	High-activity Zone Density
HQZD	High Quality Zone Density
HRZD	High-risk Zone Density
HSE	Health, Safety, and Environment
IMSI	International Mobile Subscriber Identity
KDE	Kernel Density Estimate
KML	Keyhole Markup Language

KMZ	Keyhole Markup Language Zipped
LAZD	Low-activity Zone Density
LQZD	Low Quality Zone Density
LRZD	Low-risk Zone Density
MAZD	Medium-activity Zone Density
MNT	Manual Tools
MOP	Machines Operation
MQZD	Medium Quality Zone Density
MRZD	Medium-risk Zone Density
MS	Masonry
MSP	Microsoft SharePoint
NASA	National Aeronautics and Space Administration
NND	Nearest Neighbor Distance
OOP	Object-oriented Programming
OPTICS	Ordering Points to Identify the Clustering Structure
PDA	Productivity Data Analysis
PF	Pipe Fitting
PPA	Point Pattern Analysis
PPE	Personal Protective Equipment
QC	Quality Control
QDA	Quality Data Analysis
QR	Quick Response
RA	Resting Area
RBA	Robotic Automation
RDP	Ramer-Douglas-Peucker
RFID	Radio-frequency Identification
RG	Rigging
RGSD	Real-time Geospatial Data

RGSDP	Real-time Geospatial Data Preparation
RTMS	Real-time Monitoring Systems
SC	Scaffolding
SDA	Safety Data Analysis
SE	Site Engineer
SF	Steel Fixing
SGD	Site Geographical Data
SGDP	Site Geographical Data Preparation
SL	Skilled Labor
SOCDA	Second Order Correlative Data Analysis
SP	Supervision
SPD	Site Periodic Data
SPDP	Site Periodic Data Preparation
SQB	Site Quality Behavior
SQD	Site Quality Data
SQP	Site Quality Performance
SQZ	Site Quality Zones
SSB	Site Safety Behavior
SSD	Site Safety Data
SSP	Site Safety Performance
SSRZ	Site Safety Risk Zones
TP	Travel Path
USL	Unskilled Labor
UWB	Ultra-wide Band
VSM	Value Stream Mapping
WA	Working Area
WGS84	World Geodetic System 1984
WL	Welding
IT	Information Technology
IoT	Internet of Things

n	Total Number of Observations
p	Value of Observation
p_o	Ordered List of Values of the Observations
Med	Median
μ	Mean
σ	Standard Deviation
x	x-coordinate
y	y-coordinate
x'	Median Center of x-coordinates
y'	Median Center of y-coordinates
w	Weight assigned to each Median Center
d	Distance between 2 coordinates
D	Standard Distance
W	Weight matrix
θ	Rotation Angle
μ_{NND}	Mean of Nearest Neighbor Distance
μ_{CSR}	Mean of Point in Complete Spatial Randomness
z	Standard Score
LS	Linear Sum
SS	Squared Sum
r	Radius of Cluster
T	Threshold Distance
B	Branching Factor
$K(x)$	Kernel Density Function
h	Band Width
M	Set of x and y coordinates
g	Voronoi Cell Generator
t	Time period of Analysis

S	Number of Safety Incidents
R	Number of Inspection Requests
j	Easting Coordinate
k	Northing Coordinate
C	Set of workers' coordinates
R_k	Voronoi Cell
G	G-distance Function
Gg	G-distance Envelope

DEDICATIONS

In loving memory of my grandmother, Hoda Shalaby, I would like to dedicate this research to her. She raised me and has been by my side every step of the way. I almost gave up on a lot of things after her death two years ago, including getting my master's degree. However, remembering her sweet yet firm words gave me the courage I needed to continue with this research. She was the driving force behind my decision to pursue my post-graduate studies in the first place. She always pushed me to do more and do better for myself; she always made me realize how strong I could become and that I could go out into the world and achieve whatever I set my mind to. So, if it hadn't been for her, I wouldn't be writing this today, and for that, I will be eternally grateful. Her love and drive remain alive and constant even in the happiest and most difficult of times.

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to Dr. Khaled Nassar and Dr. Khayyam Dorra for their constant guidance, support, and motivation during the highs and lows of this journey. I would also like to acknowledge Dr. Ossama Hosny for being the first professor to believe in me and give me the initial push I needed to launch my career path. In addition, I want to thank Engineer John Edwar for his dedication and aid in the completion of this research. Special thanks go to my work colleagues Engineer Mina Ebraheem and Engineer John Samir for their understanding and for adjusting our work schedule to accommodate my needs when undertaking the research. Last but not least, I would like to express the joy of having my best friend, Raghda Attia, as my companion as we both progressed with our studies.

Undoubtedly, I want to take the opportunity to show my appreciation to my loving parents, whom I deeply cherish, for their unconditional encouragement and backing throughout the different phases of my academic growth. Thank you for always pushing me when I needed it the most.

ABSTRACT

In construction, the field of real-time monitoring and control of on-site resources is still evolving. In recent years, there has been an increasing demand for improved real-time monitoring systems on construction sites. Such technologies are widely employed in domains such as IT monitoring and the healthcare industry. The construction industry, on the other hand, has yet to recognize the value of implementing real-time monitoring systems on building sites. There is limited research in the area of using real-time monitoring, such as RFID tags and Visual Sensing, for the control of on-site safety performance and labor and earthmoving equipment productivity. Many of the studies focus on analyzing the data gathered, with the information only being used to verify the accuracy of the real-time monitoring technologies employed.

The purpose of this study is to reveal the potential outcomes that can be obtained by collecting and analyzing real-time data about worker locations on a construction site. The results are obtained by creating a semi-automated framework that analyses the collected real-time data. The framework consists of three steps. To begin, site data is semi-automatically collected, and spatial data from real-time workers is obtained using GPS tracking technologies. In the second step, the data is prepared for processing. Third, to examine the prepared data, visual and spatial-temporal analysis techniques are used. After that, the proposed framework is tested on a theoretical set of data for the first order of data analysis. The framework's implementation outputs are then used for second-order analysis, where the individual outputs are compared to each other. The results are compared to identify any existing relationships between the various site parameters. The goal of second-order analysis is not to quantify or define the relationships, but rather to provide a means of comparison for identifying potential relationships.

Finally, the framework is applied for workers' spatial data collection and analysis on a case study for a construction site in Cairo, Egypt. The outputs of the framework highlight the potential benefits of deploying real-time technologies on construction sites. As a result of this research, stakeholders now have access to a framework that collects and analyses data from construction sites in order to improve site performance monitoring and control.

CHAPTER 1 – INTRODUCTION

1.1 Problem Statement

The pay-off of deploying real-time techniques to monitor on site performance parameters is yet to be realized. A project's performance could be enhanced by depicting the wastes generated on sites and implementing the necessary actions to reduce such wastes, this is known as Lean construction. Deploying the concept of Lean construction relies heavily on the accuracy, size, and frequency of the data collected. Until recent years, the data collection techniques used for Lean Construction are labor intensive . Also, the data collected was usually not exhaustive enough to identify the possible areas of wastes contributing to a projects' poor performance. Therefore, a more efficient technique to collect data in real-time would be beneficial. The data collection process would be much less labor intensive, time consuming, and would provide valuable information in real-time allowing for more enhanced data analysis. The outputs from such analysis would make it more attainable to recognize wastes, specifically, time and productivity waste, on a construction site, thus allowing decision makers to implement the necessary actions to reduce these wastes and improve projects' performance.

1.2 Lean Approach in Construction

Given the competitive nature of the construction industry, there becomes an increasing need for the optimization of the value-chain of construction projects. One of the ways to optimize the value-chain is to reduce waste. One of the major LEAN principles is the identification and minimization of waste. (Josephson & Saukkoriipi, 2005).

Waste on a construction site could occur due to a number of contributing factors (Nikakhtar et. al., 2015), some of these are:

1. *Construction Site Waste:* Waste due to wait periods, Equipment wear and tear, Resting Time, Excess materials on site, and Debris.
2. *External Factors:* Excess materials, Clarification needs, and Waste due to design errors.

3. *Construction Processes:* Over production, Safety costs, Scrap waste, Transport/handling time, Rework, Waiting time, Idle time, and Unnecessary inventories.

Given that waste generated by construction labor is a common denominator of construction waste, it became more vital to investigate the processes of activities carried out by labor on site (KO & KUO, 2015). Thus, process maps and value-stream maps were developed for the construction activities were developed to recognize value-adding and non-value-adding activities in attempts to depict process waste. Value-stream maps are similar to that shown in Figure 1-1 and process maps are similar to that shown in Figure 1-2.

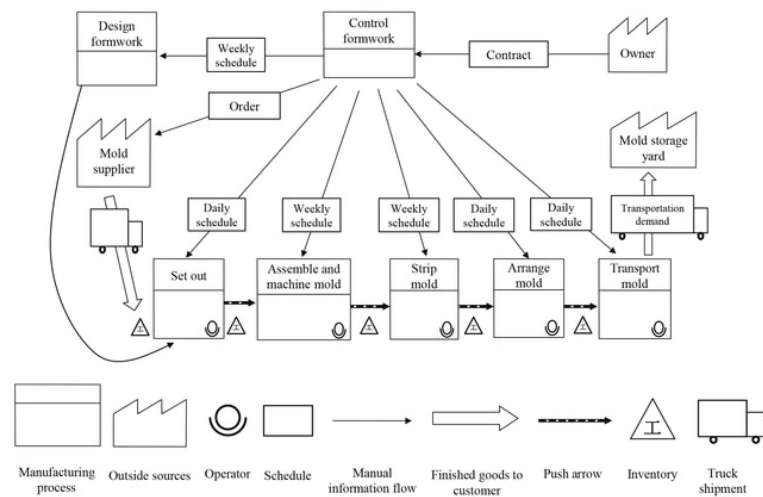


Figure 1-1: Value-stream Maps for Concrete Formwork Activity (KO & KUO, 2015)

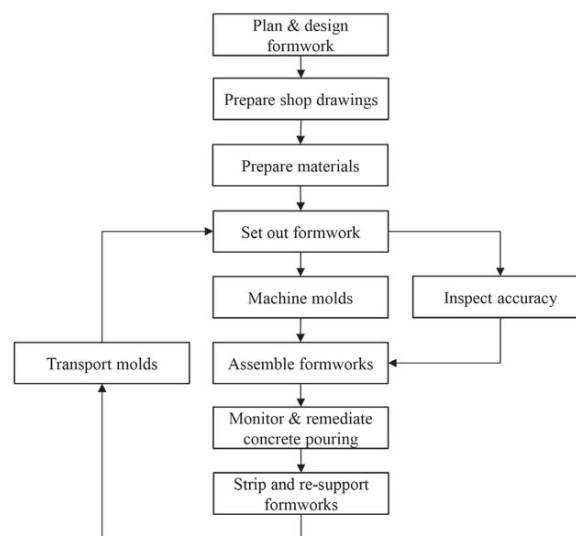


Figure 1-2: Process Map for Concrete Formwork Activity (KO & KUO, 2015)

From value-stream maps, the notion of unnecessary labor motion on site was discovered (KO & KUO, 2015). This means that there could be potential benefits of monitoring the motion of workers on a construction site could and comparing their motion to the overall productivity of the time. This helps determine the root cause of the unnecessary labor motion, whether it is an inefficient site layout or over-allocation of labor resources in a certain area. Also, the identified waste could be correlated against other site performance parameters to detect how this waste affects a projects' overall performance in terms of safety, productivity, quality, etc.

1.3 Real-time Monitoring on Construction Sites

Lean construction relies heavily on the data collected and analyzed for a construction site. Most of the data collection techniques commonly used are labor intensive, time consuming, and do not provide all of the necessary data for analysis of site performance. Nevertheless, the use of real-time monitoring techniques on construction sites has become of more interest lately. Studies conducted in this area have been limited to the application of real-time monitoring technologies in the field of health and safety. (Soltanmohammadloua et. al., 2019)

The studies showed the potential of applying technologies such as RFID tags, Bluetooth, Internet-of-Things (IoT-based), and Visualization technologies in enhancing safety performance on site. Studies have concluded that such technologies could be deployed to locate and track workers and equipment, to monitor their entrance inside a pre-determined risk zone. Also, the working environment could be monitored using wireless networks to ensure safety in high-risk construction environments. (Soltanmohammadloua et. al., 2019)

The research also looked into applying real-time monitoring technologies in measuring the productivity of workers and equipment. By using the locations of resources on a site, the resources are assumed to either be idle or engaged in the ongoing activity in the same area. The time spent in the locations would then be compared against the quantities executed during that time and the productivity would be measured. (Navon et. al.,2005)

It was concluded from the literature that a fully automated real-time monitoring system is the most effective control system. The literature also suggests that the use of GPS technologies proved to be more efficient than other real-time monitoring technologies.

However, the studies conducted in this field suggest that limited research has been done in the area of real-time monitoring of labor. However, future research must focus on this area due to its apparent effectiveness and potential that remains uncovered.

1.4 Research Objectives

The overall objective of this study is to unveil the obscure potential of using real-time monitoring technologies on construction sites for minimizing waste and enhancing project performance. The research does not quantify the means of improving site performance parameters, it rather provides a novel approach to depict areas of time and productivity waste, referred to as waste from now on.

The main objective is divided into sub-objectives as follows:

1. Collection and analysis of workers' spatial data using real-time technologies for monitoring workers' productivity on site.
2. Statistically analyzing the construction site's performance.
3. Correlating the results from the spatial analysis of workers' data and statistical analysis of site performance parameters. The correlation depicts existing patterns or relationships, if any, between workers' spatial behavior and site performance for waste and/or mismanagement identification and elimination.

1.5 Scope of Work

The scope of this research is concerned with studying the 2-D, Easting and Northing, spatial behavior of workers only. Thus, the parameter of elevation was not considered in this study. Moreover, the site performance parameters considered under this study are Safety, Quality, Productivity, Site Layout Efficiency, Site Progress, and Cost Expenditure only. Also, the research focused on the deployment of GPS tracking technologies to monitor the spatial behavior of workers on site. Finally, the monitoring device used in the developed framework was the personal smartphones of the workers.

1.6 Research Methodology

In pursuit of reaching the research objective, the research methodology was divided into three main stages.

Stage 1 – Knowledge Acquisition and Problem Identification:

1. Review of studies conducted highlighting the importance of implementing lean-construction techniques in the construction industry.
2. Detailed review and study of the literature regarding real-time monitoring of construction sites.
3. Identifying the gap in the literature regarding the implementation of real-time monitoring technologies to enhance projects' performance.

Stage 2 – Framework Development:

A framework is developed in attempts to address the problem statement. The framework was divided into 3 stages:

- Data Collection: Three types of data are collected for a construction site; (1) Site Geographic Data, (2) Site Periodic Data, and (3) Real-time Geospatial Data. (1) and (2) are collected using semi-automated techniques, whereas (3) is collected using GPS tracking technologies.
- Data Preparation: The collected data is prepared for analysis by splitting and grouping the data using Python ®.
- Data Analysis: Finally, the prepared data is analyzed using visual and spatial-temporal statistical analysis techniques. The algorithms of the techniques were applied to the data using Python ® as well.

Stage 3 – Framework Implementation:

1. The developed framework is implemented on a set of theoretical data generated for a random construction site; this is used for second-order analysis by correlating outputs from the first-order analysis.
2. The framework is also implemented on a case study to verify the applicability and benefits of using the framework on a construction site.

1.7 Thesis Organization

This thesis is comprised of 6 chapters; the contents of each chapter are listed below.

Chapter 1 – Introduction:

This chapter provides a general introduction to the topic of real-time monitoring of workers on construction sites as well as the spatial and statistical analysis that could be performed on data collected from sites. The chapter also discusses, in brief, the problem statement, research objectives, research methodology to reach the objective, and finally the thesis organization.

Chapter 2 – Literature Review:

In this chapter, previous studies conducted regarding lean construction and real-time monitoring on construction sites, are discussed in depth. Thus, highlighting the gap in literature this research aims to fill.

Chapter 3 – Framework Development for First Order Analysis:

The purpose of this chapter is to explain the development of the framework that is used to collect, prepare, and analyze spatial data collected using GPS tracking technologies on construction sites. The framework also includes the semi-automated collection, preparation, and analysis of other site parameters for further analysis.

Chapter 4 – Framework Implementation for Second Order Analysis:

Here, the framework is implemented on a set of randomly generated construction data to represent the possible outputs of the frameworks' application. Then, by comparing

and correlating the results from the first-order analysis, the outputs of the second-order analysis are achieved.

Chapter 5 – Case Study:

This chapter confers the feasibility of applying the framework on an actual construction site, by examining the results achieved from such application.

Chapter 6 – Conclusion and Recommendations:

The research is concluded and the recommendations for future developments in the area of real-time monitoring on construction sites and spatial analysis of real-time data are provided.

CHAPTER 2 – LITERATURE REVIEW

2.1 Waste Allocation (Lean Construction)

Part of the optimization of the project's performance measures, is being able to identify and minimize waste in the process of construction, with waste in the construction process accounting for 30% to 35% of the project's total cost (Josephson & Saukkoriipi, 2005). The identification and minimization of waste are one of the major LEAN principles, and according to these principles, waste could be divided into 4 main categories as follows:

- Defects and checks: These are wasteful costs spent on managing defects, including the costs of inspection, insurance, theft, and destruction.
- Use of resources: This covers the unnecessary costs associated with an inefficient use of labor, machines, and materials, this category accounts for more than 10% of the project's cost.
- Health & Safety: This is the extra cost related to work-related injuries and illnesses.
- Systems and structures: These are the cost of waste in relevance to long land-use planning processes, documentation, inefficient purchasing processes.

A study conducted by Thomas et al. (2003) investigated the use of lean principles in optimizing the workflow of labors to minimize the associated wastes and lead to better labor performance. The study was conducted by collecting data from three bridge construction projects. Data collection included data about the working hours, quantities executed, and required quantities according to the schedule, to be able to calculate the rate of the inefficiency of labor, i.e., wasteful hours. The research discussed the importance of labor flow, which involves tracking the movement, allocation, and interaction of labor crews to the different ongoing activities. It is taken into consideration that in construction the number of required manpower on-site is dynamic, depending on the nature of the ongoing activities, schedule demands, design errors and changes, weather, and work sequence. The highly dynamic nature of labor flow in construction makes the need for its management even more crucial to the performance of the project. Finally, results concluded that there were major deficiencies in utilizing labor resources, where 58% of inefficient labor workhours were due to ineffective

management of workforce flow, meaning that there should be more focus on implementing lean principles on workforce management strategies to enhance project performance (Thomas et. al., 2003).

Previously mentioned studies, then lead to the further categorization of possible construction waste that needs to be managed, where waste was then categorized according to the Lean approach as shown in Figure 2-1. (Nikakhtar et. al., 2015)

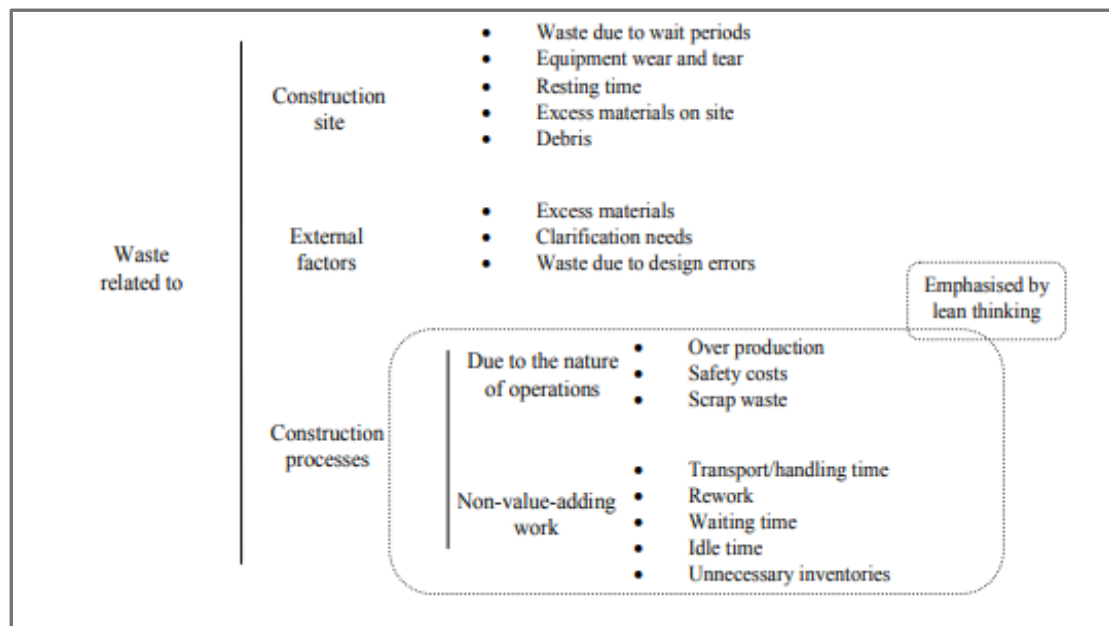


Figure 2-1: Waste Generated due to the Nature of Operations (Nikakhtar et. al., 2015)

Also, the need for managing the workflow of the workforce became more crucial, as research suggested that the first lean principle to minimize waste is to manage the flow of processes, inherently managing the workforce by attempting to minimize unnecessary motion, queue, interruption, waiting time of workforce, to maximize the efficiency of the available manpower. The mentioned Lean approach was then applied to one of the processes, carried out in a real-life construction project, which is the rebar process. Originally in the project, for every rebar workstation, all the rebars are delivered as a single batch to the next workstation, i.e., all the rebar is first cut then are delivered to the bending station and so on until the process is complete. A process map of the rebar activity, shown in Figure 2-4, was used to determine value-adding and non-value-adding activities. It was noticed that the flow of this specific process was not very efficient and resulted in a lot of waste, majorly waiting time, idle time, rework time, and full inventory. To apply the flow management technique, instead of having full batches of rebar delivered from one station to the next, the full batch was broken into

sequential smaller size batches, this allows for the concurrency of work of the different workstation and thus reduces the waiting and idle time of the workforce as shown in Figure 2-2. (Nikakhtar et. al., 2015)

As a subsequent result, the workforce was re-allocated to speed up the cycle time of the process, proving that initially the allocation of the manpower was inefficiently utilized. The reallocation is as shown in Figure 2-3. (Nikakhtar et. al., 2015)

The implementation of the Lean model showed a 9.22% improvement in the cycle time of the reinforcement process, confirming the improvement of processes through implementing Lean principles in construction. (Nikakhtar et. al., 2015)

<i>Type of waste (in each cycle)</i>	<i>Real-world model</i>	<i>Lean model</i>	<i>Improvement (%)</i>
Rework (man-hours)	0.47	0.08	82.98
Waiting (man-hours)	27.04	23.73	12.24
Transporting/handling (man-hours)	1.72	1.55	9.88
Unnecessary inventories (number of rebar pieces)	20.38	5.14	74.78

Figure 2-2: Process Waste Comparison between the Real-world and Lean Model (Nikakhtar et. al., 2015)

<i>Activity</i>	<i>Resource</i>	
	<i>Real-world model</i>	<i>Lean model</i>
Hauling four rebars	Labourer1, Labourer2	Labourer1
Cutting rebars into three pieces	Steel Fixer1, Steel Fixer2	Steel Fixer1
Cutting inspection	-	Foreman
Rework for cutting	-	Steel Fixer1
Hauling 12 pieces of cut rebars to bending area	Labourer1	Labourer1
Bending 12 pieces of rebars	Steel Fixer1, Steel Fixer2	Steel Fixer2
Bending inspection	-	Foreman
Rework for bending	-	Steel Fixer2
Hauling 24 pieces of rebars to working area	Labourer1, Labourer2	Labourer2
Rework	Foreman, Steel Fixer1	-
Placing the rebars	Foreman, Steel Fixer1	Foreman, Steel Fixer1
Tightening	Foreman, Steel Fixer2	Foreman, Steel Fixer2
Placing the spacers	Foreman	Foreman

Figure 2-3: Resources Assigned to Activities in the Reinforcement Operation (Nikakhtar et. al., 2015)

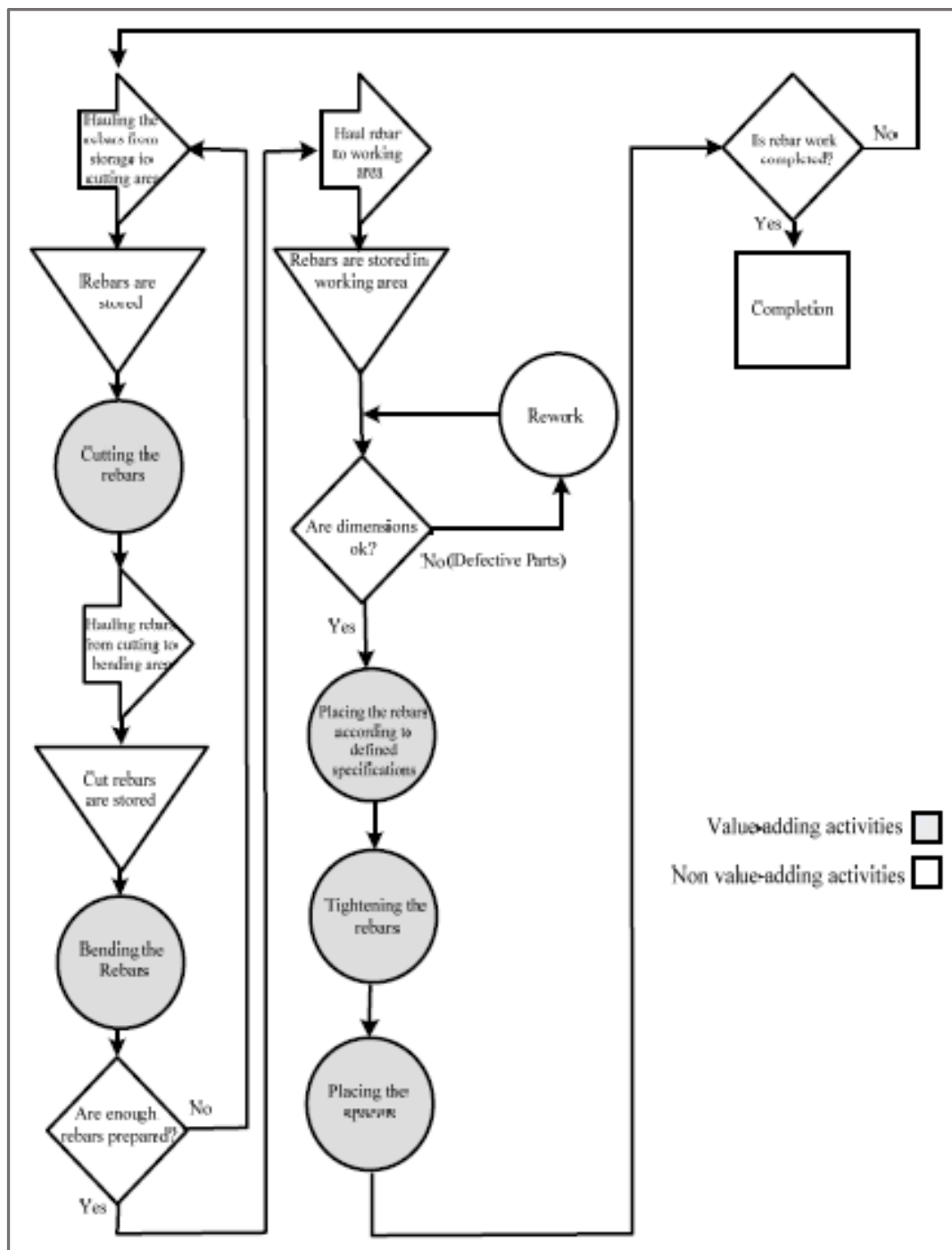


Figure 2-4: Rebar Activity Process Map (Nikakhtar et. al., 2015)

Another implementation of Lean principles in process improvement was conducted by Wang et al. (2015), where research used these principles alongside simulation optimization to address the efficiency of the layout design and staff assignment for a combined hospital Emergency Department (ED) in Taiwan. The study utilizes the principle of Value Stream Mapping (VSM) to investigate the problems mentioned above. After which, simulation optimization was used to optimize the staff assignment, the level of service, and the waiting time of patients in the department. VSM was carried out for the original state of the layout of the department and after the simulation of the different possible layouts as shown in Figure 2-5.

	Waiting time (minute)	Service level (%)	Staffing level reduction
Original system	78	54.86	9
Optimal solution	38	88.55	6
Improvement (%)	51	61.41	33.33

Figure 2-5: VSM Output (Wang et al., 2015)

Thus, using VSM, a noticeable improvement was documented in the entire process of patient care in the ED studied, highlighting the promising potential of using Lean principles to optimize workflow processes.

2.2 Real-time Monitoring Systems - (RTMSs)

Real-time monitoring on a construction site has become more popular as it has made it easier to control the productivity and safety of labor and equipment. Research has focused on attempts to fully automate the project performance control, as well as calculating performance indicators from indirect parameters that are sought after using real-time monitoring. Navon (2015) highlighted the use of real-time monitoring on a construction site by focusing on certain project performance indicators, as a measure of a project's success, such as:

1. The productivity of labor and earthmoving equipment is based on their location at regular time intervals.
2. Progress based on the location of labor and equipment or using data collected from tower cranes.
3. Full-time monitoring of materials from the moment they arrive at the site till they are used on site.

4. Prevention of accidents due to working at height by monitoring the status of guardrails.

2.2.1 RTMS Application for Health & Safety

One of the common utilizations of RTMSs in the Construction Industry is the application of the system in the field of health and safety. A study conducted by Nazi Soltanmohammadloua et. al. (2019) provides a comparative review of the most predominant research done in the application of RTMSs in monitoring labor and equipment for better site safety. The study states that despite the significance of utilizing RTMSs for enhanced safety on-site, research in the area has been limited and there seems to be a gap in the literature regarding this topic. However, the work done in RTMSs has been able to aid safety management processes in the eight major research streams mentioned below:

1. Safety Monitoring
2. Accident Prevention
3. Behavior-based Safety
4. Safety Alerts and Warnings
5. Ergonomics Analysis
6. Physiological Status Monitoring
7. Communication-based Safety
8. Performance Evaluation of the Developed RTMS-related Technologies and On-Site Safety Training

Results from the application of RTMSs in the above eight streams have been highlighted, however, only those of the first three streams will be discussed.

1. Safety Monitoring

- a. Safety monitoring of workers: RTMSs were used to locate and track construction workers using RFID, Bluetooth, Internet-of-Things (IoT-based), and Visualization technologies. The technologies helped provide major data for visualizing workers in compromised safety situations, specifically in confined spaces, working at heights, back over accidents, and in detecting workers' unsafe behaviors.

- b. Safety monitoring of equipment: RTMSs have mainly been used for crane safety using RFID, GPS, Wireless Sensor Network, and Visualization technologies. The major benefits of this application are that it collects Jobsite data, improves crane operator visibility, and recognizes the entrance of workers inside a pre-determined risk area where heavy equipment is being operated. Also, detection and tracking of excavator operation and movement of heavy objects during hoisting has made it easier to support operators' safety in offshore sites.
- c. Safety monitoring of equipment & workers: Similar technologies, to those mentioned before, have been used to monitor the simultaneous locations of equipment and workers on-site making it possible to prevent unauthorized entries to pre-determined hazardous zones, detect the proximity of dynamic objects to one another and track the movement of temporary resources in infrastructure projects.
- d. Safety monitoring of working environment: In this application, wireless sensor networks were used to attain onsite information in real-time such as temperature, humidity, and hazardous gas levels especially in high-risk construction environments.

2. RTMS technologies-aided accident prevention

Also, RTMS technologies have been used for real-time collision accident prevention by, for example, increasing awareness of operators of on-site real-time collected data, such as working areas on-site and dynamic movement of any objects in the surrounding area by providing real-time locations of workers and other equipment. Thus, reducing the risk of collision-based accidents, struck by falling object accidents, and near-miss events.

3. RTMS technologies-aided behavior-based safety

Behavior-Based Safety (BBS) aims to estimate the safety performance of on-site personnel by assessing the rate of safe behavior exhibited by such personnel. RTMSs can assist BBS by providing real-time warnings and detecting any unsafe behavior by highlighting the prevalent posture of personnel and misuse of Personal Protective Equipment (PPE). It can also be applied for path safety by collecting the trajectory data of workers on site. (Soltanmohammadloua et. al., 2019)

The study presented concludes that although RTMSs have proven to have potential in providing safety management with an automated real-time system that detects and tracks unsafe behaviors of workers and unexpected hazardous operation of heavy equipment, nevertheless research in this area is insufficient and non-comprehensive. The research done lacks to cover safety-related factors such as workers' motion and posture, it has also been limited to certain types of accidents and does not consider the application of RTMSs in heavy construction projects or indoor construction environments. Another major lack in research is that of providing a long-term assessment of the efficiency and effectiveness of deploying various RTMSs to maintain health and safety on-site. Thus, the study recommends the in-depth application of RTMSs in various other health and safety aspects should be done to enrich this research gap.

Most recently, Zhang et. al. (2021) tried to fill the research gap highlighted by Soltanmohammadloua et. al, by conducting research that develops a real-time analysis and early-warning safety system for deep large underground caverns. This system generates automatic early-warning levels and corresponding emergency plans. The early-warning system relied heavily on sensing deformations of the surrounding rock masses and the stress state of the support structures and accordingly carry-out alert-situation analysis reflecting the safety status of underground conditions and generating an early warning system.

The system was comprised of a Remote Wireless Transmission Model which had 3 layers: the sensing layer, the transmitting layer, and the application layer shown in Figure 2-6.

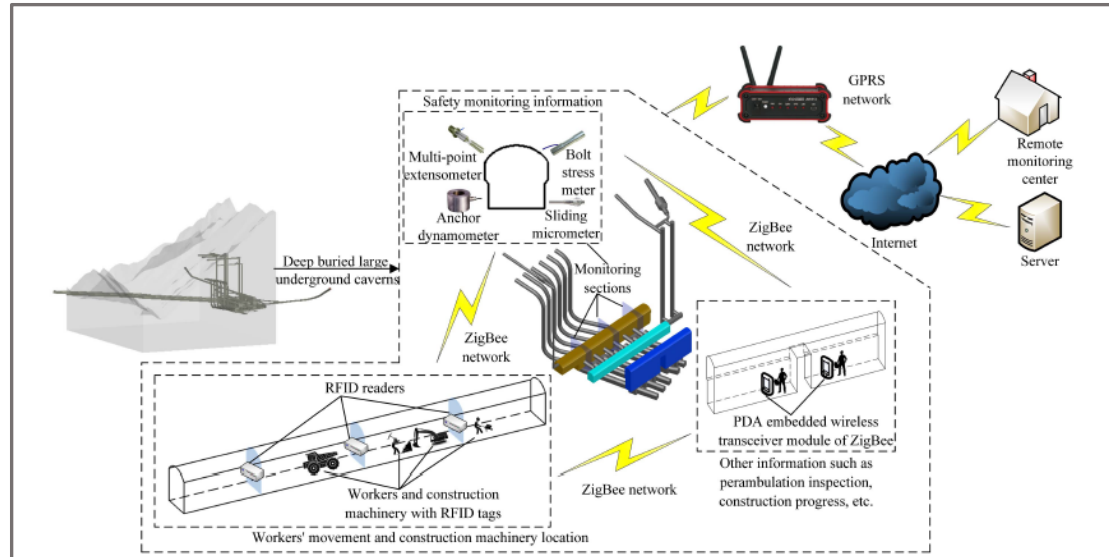


Figure 2-6: Layout of an IoT-based wireless monitoring platform for deep large underground caverns (Zhang et. al., 2021)

For the sensing layer, wireless and real-time data collection methods were used, these were:

1. Sensor-based Acquisition: Stress meters and extensometers were used to collect data regarding the deformation of the surrounding rock during the construction of the caverns.
2. Location-based Acquisition: RFID tags and readers were utilized to collect data about the workers' and Equipment's movement and location within the underground caverns; and
3. PDA-type Acquisition: using PDA embedded wireless transceiver module of ZigBee technology made it possible to collect information that the sensors were unable to perceive such as construction progress, perambulation inspection, excavation, and support, and provided real-time communication between the site and the control center.

The transmitting layer is a crucial link for the system, as it connects the sensing layer to the application layer. The layer uploads data collected by ZigBee, from the sensing layer, to the internet through the mobile base station. Figure 2-7 shows the integration of the abovementioned technologies in the remote wireless transmission model. (Zhang et. al., 2021)

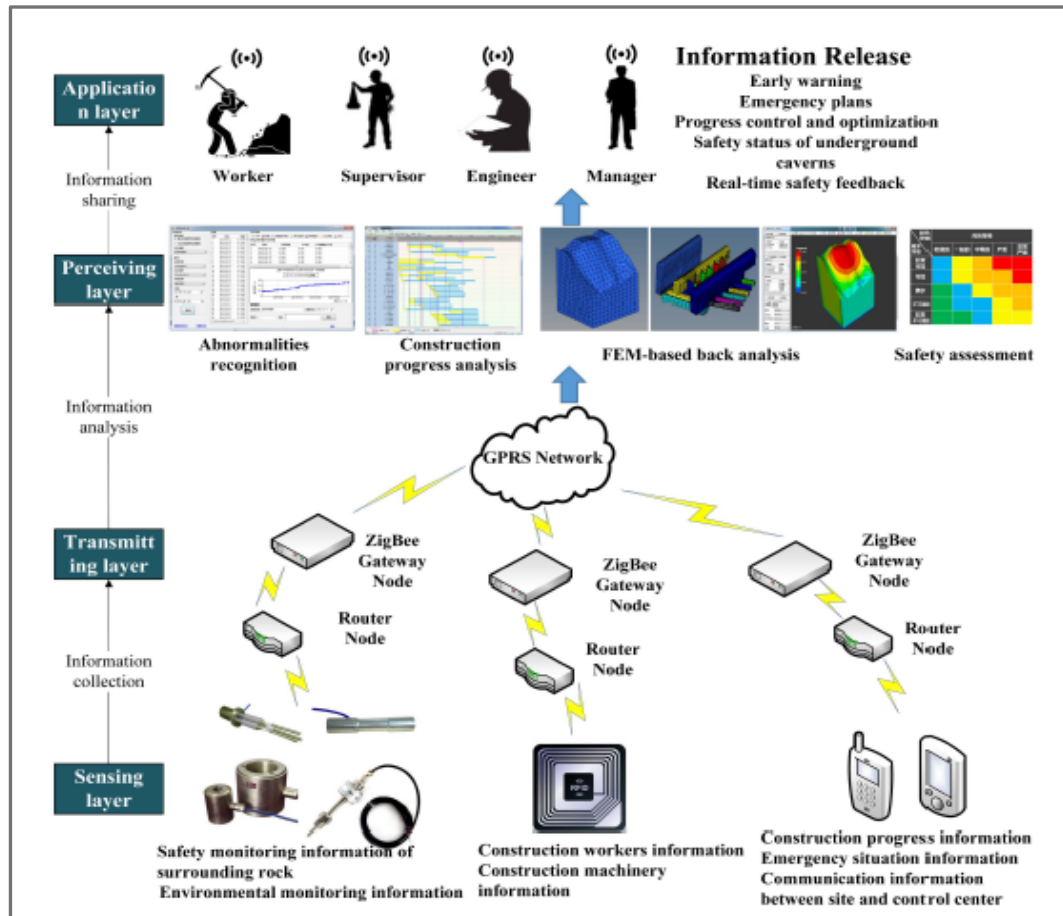


Figure 2-7: Remote Wireless Transmission Model (Zhang et. al., 2021)

Finally, for the application layer, this is the highest level of the architecture of the wireless monitoring system, where the main control center host receives the data and proceeds to perform data parsing restoration, stores the data in the monitoring database, and then uses an expert information system to perform numerous analyses and process the real-time data in the database. The architecture of the Iot-based wireless monitoring system is shown in Figure 2-9 and the architecture of the real-time and online analysis and early-warning safety system is shown in Figure 2-8.

The system developed was successfully implemented in an underground cavern construction project for Lianghekou Hydropower Station in China. The successful implementation indicates that the proposed system can recognize abnormalities from real-time monitoring of data and the evaluation process of the data makes the early-warning layer more accurate, efficient, and reliable. Hence, the application of real-time monitoring in the field of safety has proven to help identify more potential accidents and prevent their occurrence. (Zhang et. al., 2021)

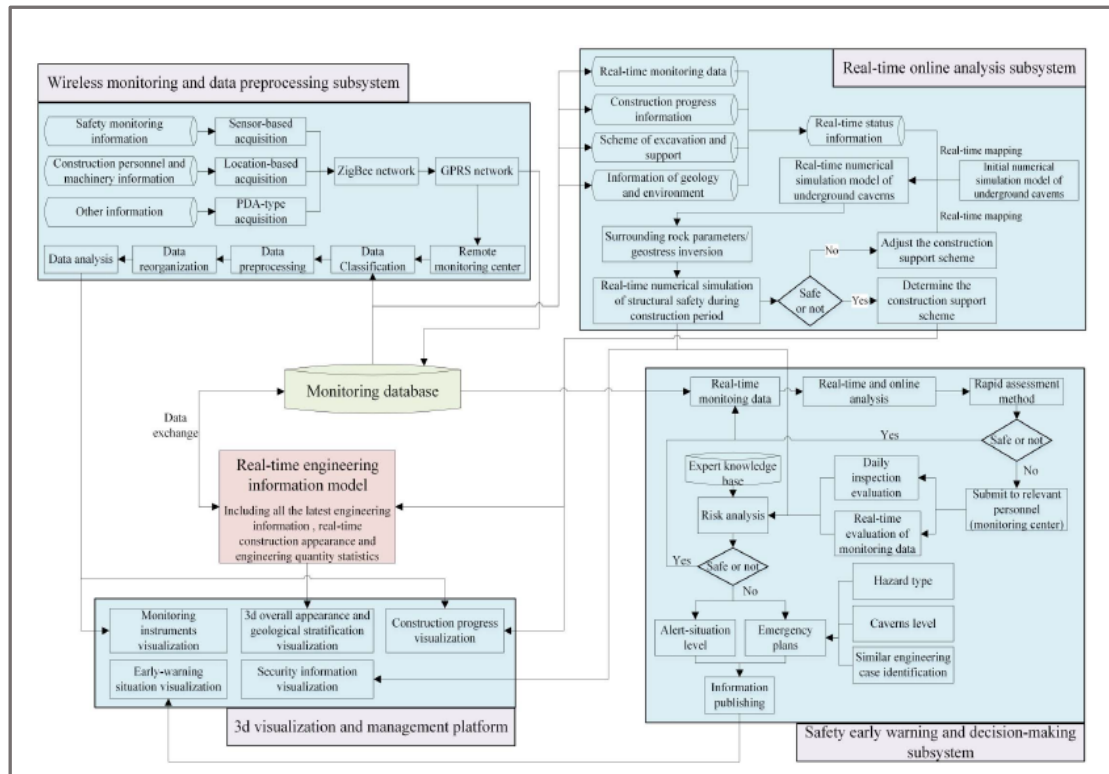


Figure 2-8: Architecture of the real-time and online analysis and early-warning safety system (Zhang et. al., 2021)

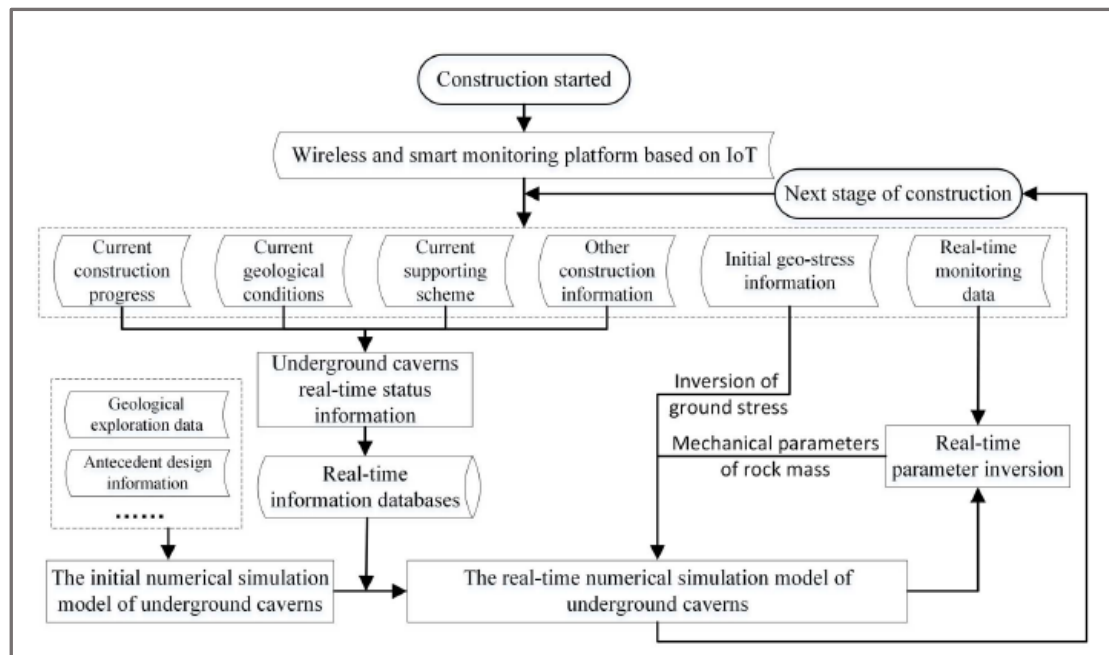


Figure 2-9: Architecture of the IoT-based Wireless Monitoring System in Underground Caverns (Zhang et. al., 2021)

2.2.2 RTMS Application for Workers' and Equipment's Productivity

Another field in which RTMSs application has been studied, is the field of on-site productivity. In 2005, Navon et. al. conducted a research exploring the deployment of RTMS technologies in measuring productivity. The research recommends the use of automated monitoring of labor and equipment versus previously deployed partially automated control methods, such as RFID, Barcodes, or PDA technologies, that relied heavily on the input of labor and thus still requiring manual work. The automated system is based on collecting data using Global Positioning Systems (GPS), with the assumption that the presence of labor or equipment near the area of work meant that the resource was productive.

Moreover, Navon and Goldschmidt (2003) had worked on a similar concept of using the location of labor/equipment as an indirect measure of their productivity. Essentially, the presence of labor/equipment at a certain time in a certain location was an indication of their engagement in the on-going activity at the time. The location-to-activity relation meant that initially a work envelope had to be developed, breaking down the working areas according to the nature of the activities taking place on site. A second relation was based on an algorithm known as logical association, which uses work continuity,

statistical considerations, or crew affiliation to link the location of labor/equipment to the type of work taking place.

After locating labor using GPS, the time the labor spent in a certain working location was measured against the amount of work performed i.e., the quantities executed during that time, to measure the productivity of the labor. The model that was developed using the concept of location-to-activity relation was implemented on a construction site, with the results from the model compared against those calculated manually, with only 12% difference noted. (Navon & Goldschmidt, 2003)

Both Navon and Goldschmidt (2003) and Navon (2005) concluded that a fully automated monitoring system based on the location of resources using GPS technologies proved to be the most effective control system as it a real-time data collection system. However, the studies highlighted that limited work has been done in real-time monitoring of labor, and more focus on this area should be put due to its apparent effectiveness.

Teizer (2015) proposed the use of an existing simple technology in videoing as a new monitoring system and technology that is vision-based sensing and tracking of temporary resources. This system involves translating real-time images into real-time information by using a camera or video-based monitoring technologies combined with processing algorithms. The research conducted states that it is possible to track the resources on site using time-lapsing, as it documents the daily workflow of on-site activities, providing important information about site-activities for more informed decision-making processes. The time-lapsed images are then connected to an algorithm that includes the schedule of the project, the site-layout, the site-coordinates, and the Building Information Models (BIM) of the site. These are used to create the geospatial link between the images and the activities on site. This model could then be utilized for progress and productivity measurements by documenting the quantities executed and the working hours of labor/equipment on site. The conceptual framework for integrating time-lapsing/videoing, suggested by the research, is shown in Figure 2-10. (Teizer, 2015)

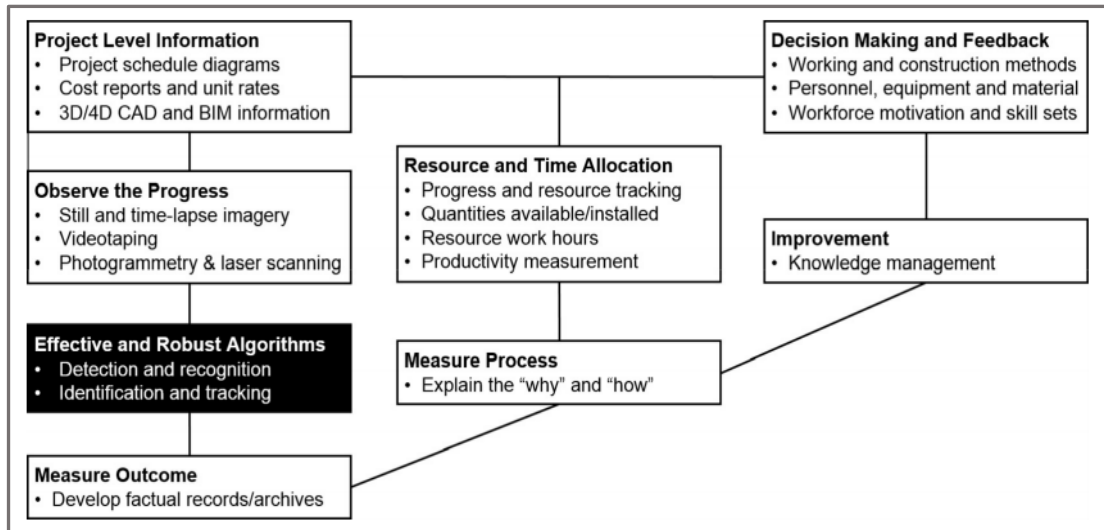


Figure 2-10: Conceptual Framework for Integrating Time-lapsing or Videoing (Teizer, 2015)

Moreover, the need for real-time data on a construction site was highlighted by identifying the four dimensions of data in a project which are, volume, velocity, variety, and value. The size of data, the speed of gathering data, the different types of data, and the necessity of the data are all major aspects to consider regarding data in a project upon which all knowledge-based decisions are made, as shown in Figure 2-11. Therefore, it becomes a major challenge where all four dimensions need to be satisfied for better monitoring and more enhanced project performance, whereas by integrating the mentioned technology, it becomes less of a challenge. (Teizer, 2015)

In the same study conducted by Teizer (2015), video-based sensing and tracking has been used mainly to monitor the following: construction personnel, large machinery, presence of containers, change in construction site layout and roads, lay down areas, supportive structures like fencing and guardrails, as shown in Figure 2-12. However, in monitoring construction personnel, moderate success has been noticed using this system. As the visual foot-print intended is quite large, the equipment used does not provide enough data when it comes to tracking smaller bodies as the resolution decreases for further distances. Therefore, the use of this system for long-term tracking is limited and might require major modifications to its tracking algorithms, as the

appearance model of an object does remain constant over-time as the imaging conditions change over-time.

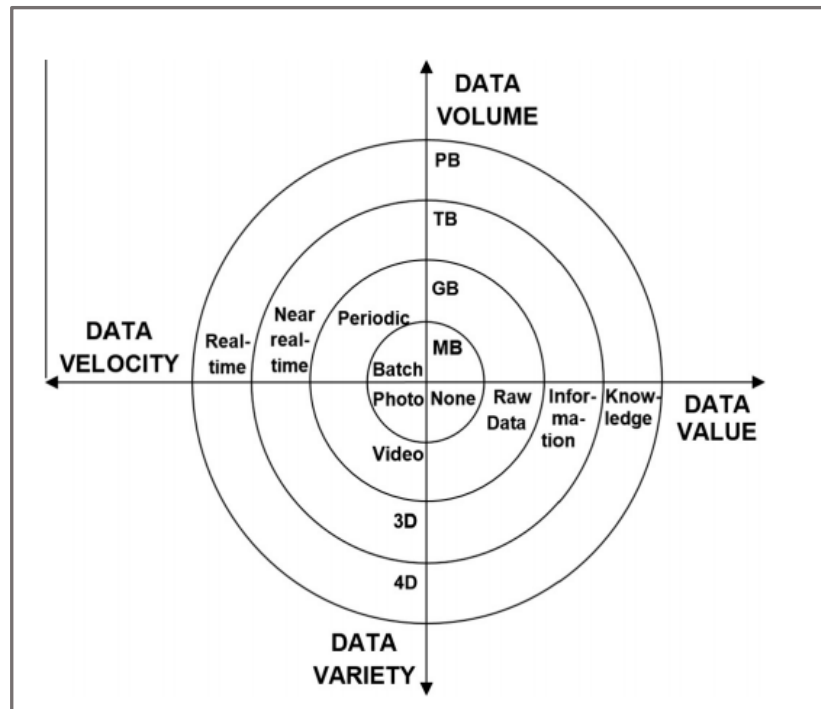


Figure 2-11: Data Aspects to Consider in Knowledge based Decision Making (Teizer, 2015)

Nevertheless, although the system may be implemented successfully in some cases, there are still major open challenges to the field of video-based sensing, as firstly the construction site does not resemble a laboratory like environment which is optimal for utilizing video-sensing technologies. Moreover, significant expertise regarding camera

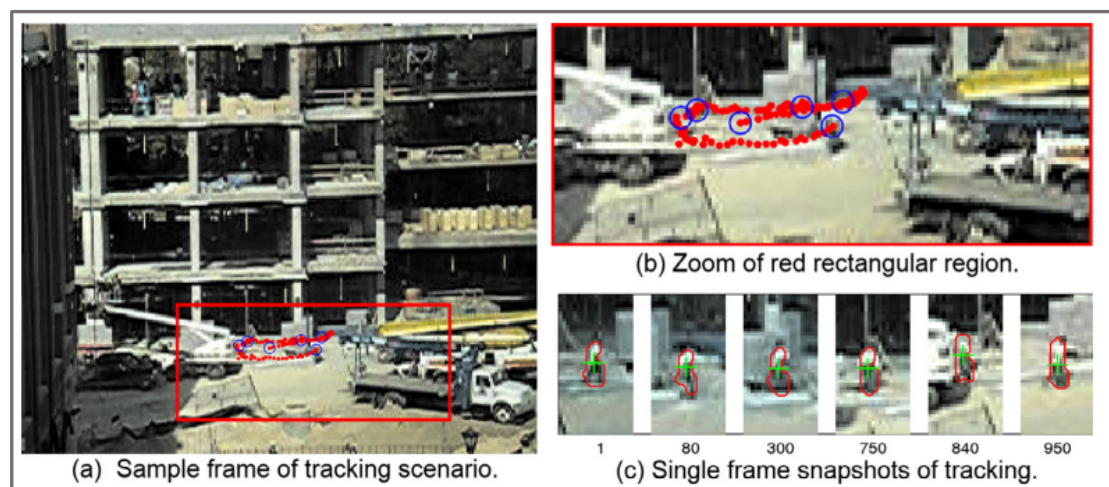


Figure 2-12: Output of Visual Tracking (Teizer, 2015)

technology is crucial for the design and testing of the proposed algorithm. (Teizer, 2015)

Another study conducted by Jiang et al. (2015) focused mainly on labor monitoring by using Global Positioning System (GPS) or Geographic Information System (GIS). The tracking technology used was implemented using smart-phones, servers, and on-site wireless base stations, display, and application. The analysis from the system could then be used to provide accurate information for decision making especially when negotiating payments with contractors. The system was firstly implemented to the third largest hydropower plant (dam) project in the world, where labors' locations and working hours were constantly monitored in real time. The framework for the system is shown in Figure 2-13.

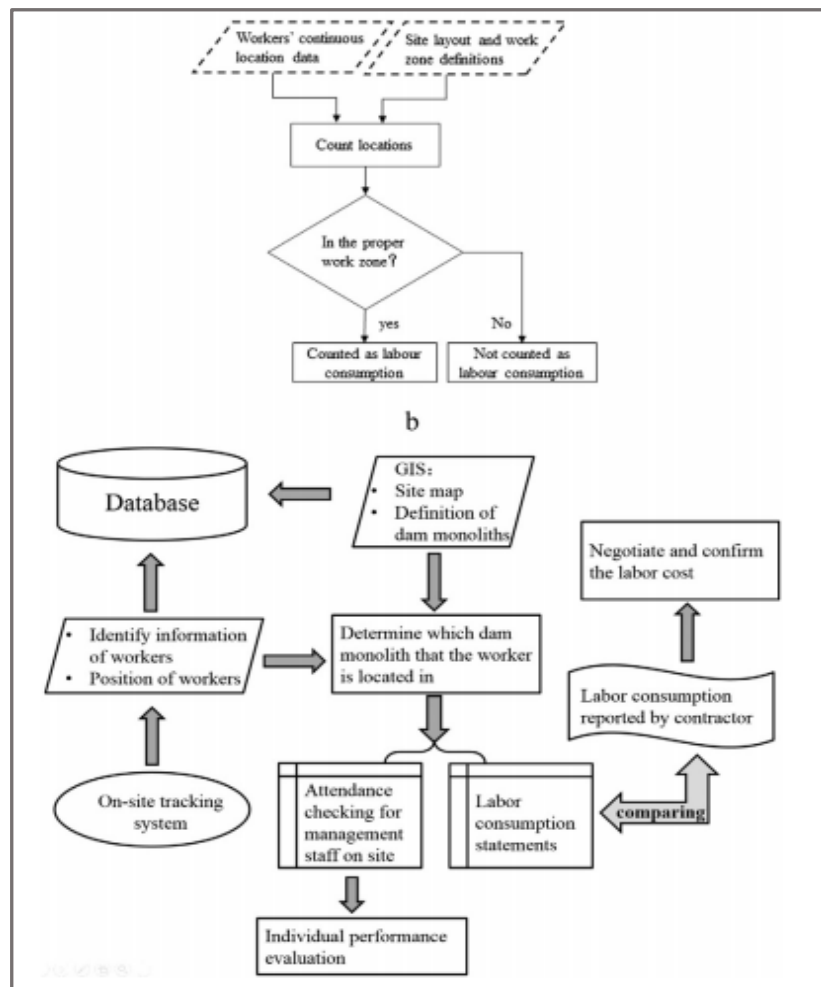


Figure 2-13: Real-time Monitoring System Framework (Jiang et al., 2015)

The successful implementation of the system presented above required the following:

1. Real-time trajectories of on-site labor.
2. Continuous registration of labor time periods spent working on each dam segment.
3. GIS data that is constantly updated according to the boundary changes of the segments.
4. Documentation and storage of quantities executed by labors for all time periods and labors' attendance records.

The system proposed consisted of three layers:

1. Data acquisition and transmission using ZTE smartphones with a built-in GPS module and private 3G network base stations, Figure 2-14.



Figure 2-14: ZTE Phone Application Interface (Jiang et al., 2015)

2. Data storage and processing, this layer (hardware) consists of 3 parts:
 - Location database: This database stores the location data of each smartphone, the smartphones are registered in the system prior to use. Each smartphone has an ID based on the name, position, unit, type of work, and IMSI (International Mobile Subscriber Identity). Depending on the IMSI, the system collects data regarding geographic coordinates, and timestamps.
 - GIS server: The server (built using ArcGIS Server and C++) provides the coordinates of the site layout and the different dam segments.

- Web server: This server (built using JAVA, C++, and Linux) uses a Socket [56] to support exchange of real-time data between the smart phones and the location database server.
3. Data display: Finally, the last layer consists of a user interface to display the collected data and view the labor consumption statistics, in Figure 2-15.

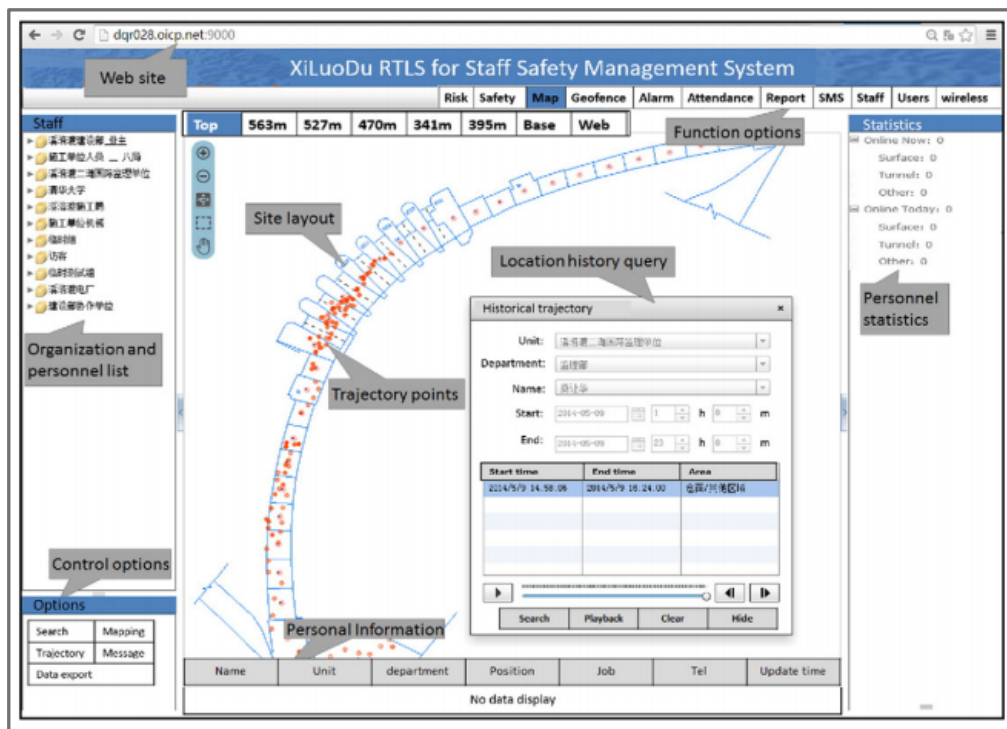


Figure 2-15: Web-based Management Application Output Interface (Jiang et al., 2015)

After the system was implemented and tested, results concluded that system was successful in gathering information with in an error that is no more than 7%, from the actual quantities registered on site, which is considered acceptable for most project managers. The system also proved to be feasible and effective as it provides a fully automated management system for on-site labor consumption, which its management required a great deal of effort and is a significant contributor to the success of the project. (Jiang et al., 2015)

The construction industry in the near future is expected to heavily invest in three top technologies, BIM dimensions analysis models, sensing technology, and business information models. Given that the construction industry is being geared towards digitalization, Calvetti et. al. (2020), highlights that there is a gap in monitoring workers' productivity, thus, the research aimed to further explore the potential

implementation of technologies for near real-time monitoring for measuring and modeling workers' productivity on site.

Calvetti et. al. (2020) was able to develop a systemized framework that measures workers' productivity based on their motion. The systemized framework, Worker 4.0, integrates nine processes on a flowchart to streamline task processes assessment and mechanization level.

Finally, the output of the framework is meant to provide a tool that aids different stakeholders in focusing on improving skills, efficiency, mechanization, and productivity. However, the implementation of sensing technologies can be highly complex and may require specialized resources. The sensing technologies were broken down into portable and/or wearable devices among others:

1. RFID (radio-frequency identification)
2. UWB (ultra-wide band)
3. GPS (global positioning system)
4. Bar code/QR Code, labels, or passive tags
5. Smartphone devices

Other technologies are listed in Figure 2-16 below as well as the data that could be gathered by using these technologies in Figure 2-17. (Calvetti et al.,2020)

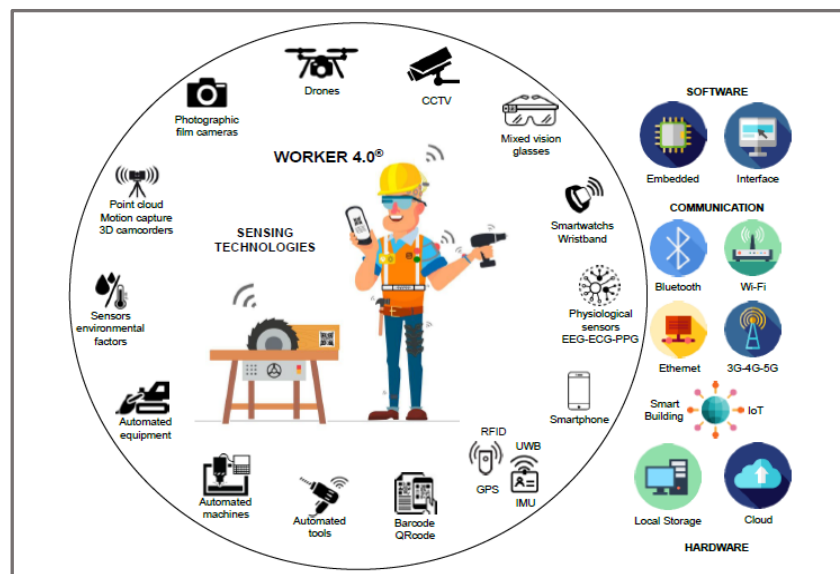


Figure 2-16: Sensing Technologies (Calvetti et al.,2020)

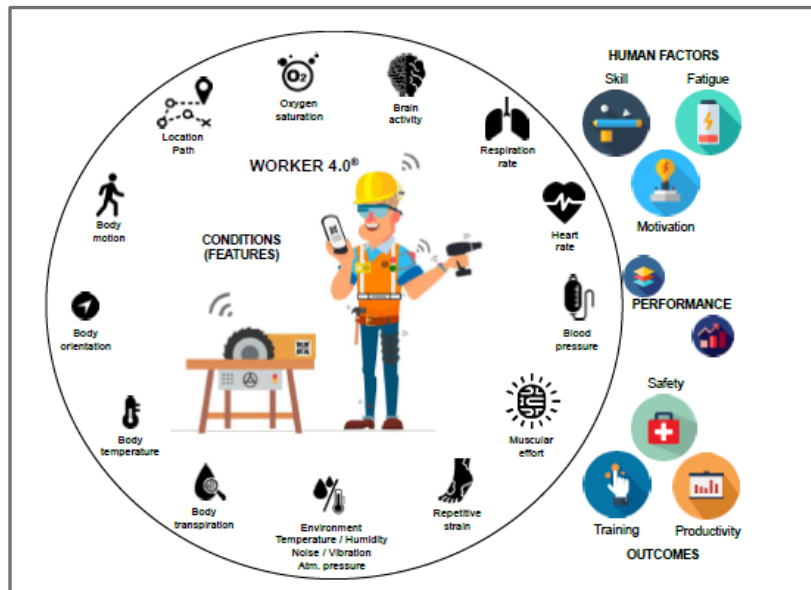


Figure 2-17: Data Gathered by Sensing Technologies (Calvetti et al.,2020)

Based on the above, the framework proposes the mixing of work elements (such as, placing a brick, using a drill, etc.) and basic motion elements (such as, walking, carrying, free hand performing, etc.) to map the processes of all construction tasks carried out on site. In addition, the study sets common basis for measuring the energy spent on each process by the workers. Figure 2-18, shows the nine basic processes of workers' motion productivity and their acronym s detailing the following:

1. Task element level (work element or basic motion element)
2. Processes Characteristics (operation, inspection, delay, transportation/storage),
3. Productivity State (productive or direct work, contributory or support work, nonproduction work)
4. Electronic Monitoring (body motion, location, sound/noise)

Additionally, the developed framework proposes a flowchart that could enhance workers' productivity as shown in Figure 2-19. (Calvetti et al.,2020)

Worker 4.0 Motion Productivity	Acronym	Element Level of the Tasks	Process		Productive State	Monitoring
			Indication	Symbol		
Free-hand Performing	FHP	work element	Operation	O	Productive or Direct work	BM
Auxiliary tools	AUT	work element	Inspection	□	Contributory or Support work	BM
Manual tools	MNT	work element	Operation	O	Productive or Direct work	BM
Electric/electronic tools	EET	work element	Operation	O	Productive or Direct work	BM + So
Machines operation	MOP	work element	Operation	O	Productive or Direct work	BM + Loc + So
Robotic automation	RBA	autonomous	Operation	O	Autonomous Production	BM + Loc + So
Do not operate value	IDL	basic motion element	Delay	D	Nonproduction work	BM + Loc
Walking	WLK	basic motion element	Delay	D	Nonproduction work	BM + Loc
Carrying	CAR	basic motion element	Transportation/Storage	$\Delta \rightarrow$	Contributory or Support work	BM + Loc

BM—body motion; Loc—location (x; y; z); So—sound detector.

Figure 2-18: Motion Productivity Specifications and Patterns (Calvetti et al.,2020)

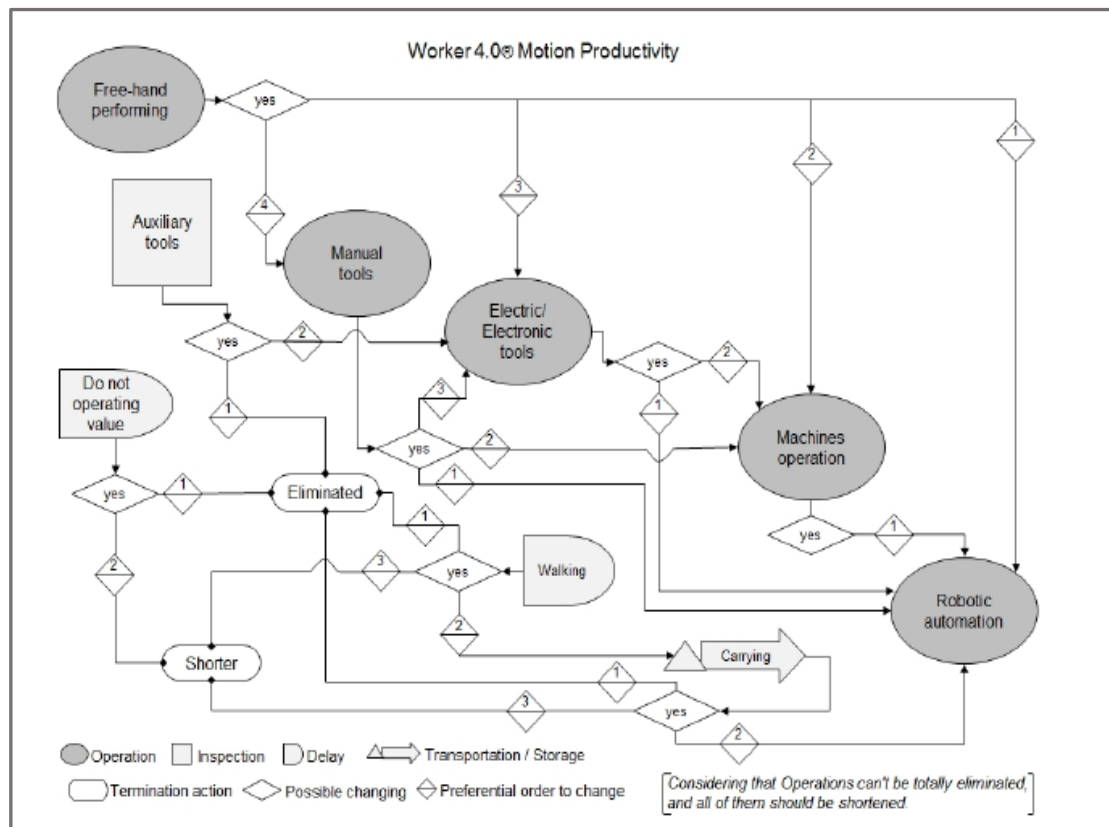


Figure 2-19: Flowchart to Increase Efficiency (Calvetti et al.,2020)

The study mentions that until today there is no academic research in the area of assessing the mechanization in the Construction Industry. Hence, the framework incorporates a mechanization index assessment tool using the Equation below. (Calvetti et al.,2020)

$$\text{Mechanization Index} = \frac{PEET + MOP + RBA}{\sum \text{Operation Processes}} \quad (1)$$

Where:

EET = electric/electronic tools;

MOP = machines operation;

RBA = robotic automation;

Operation process = (FHP + MNT + EET + MOP + RBA);

FHP = free-hand performing;

MNT = manual tools.

The analysis of these results of building construction tasks is based on the classification scale proposed:

- Low — 0% to 20%;
- Moderate — 21% to 40%;
- High — 41% to 60%;
- Very high — above 61%.

Finally, Figure 2-20 presents the results obtained from implementing the framework.

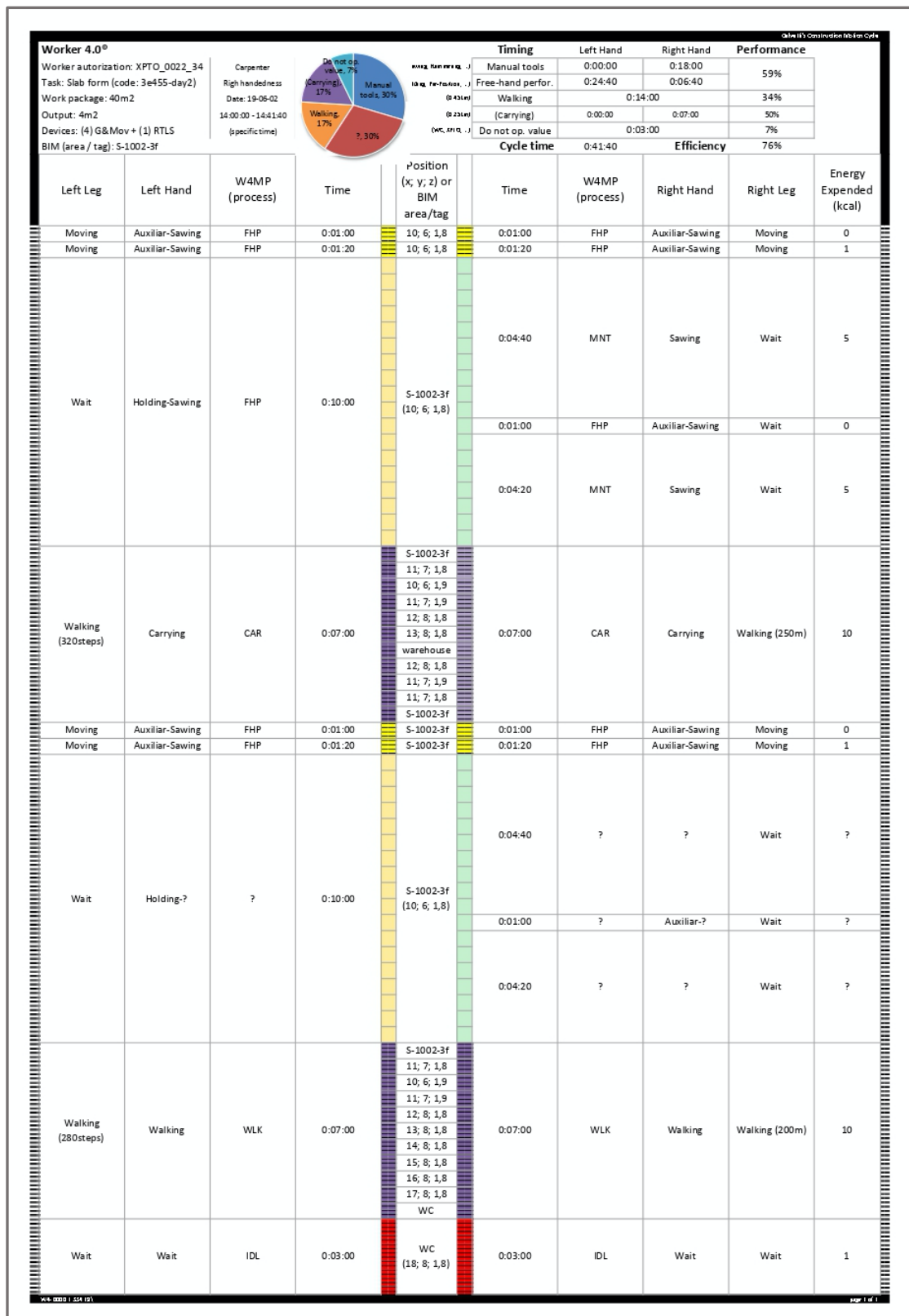


Figure 2-20: Productivity Modelling Chart (Calvetti et al.,2020)

The research then finally discusses the practical implementation of the proposed mechanization system in Construction firms stating that, the top management of the company must authorize the implementation of mechanized monitoring systems and the company shall define a clear plan to implement such systems, as well as installing the sensors. It is emphasized that workers being monitored should receive proper training in how to use sensing devices, and their consent is condition precedent to implementing the technology. Also, other stakeholders should be aware of the system as they will be interacting with it, those are:

1. Directors
2. Managers
3. Field Engineers
4. Human Resource Specialist
5. Planners and Quality Engineers
6. Field Workers

All in all, the framework proved that it could potentially have a direct impact on the works performed, workers' behaviors, and site conditions. However, there remain challenges and gaps in its implementation to date and its added value could be improved through integrating measurement methodologies and data collection devices. Accordingly, the research states that future studies should focus on conducting a set of laboratory/on site experiments that implement electronic monitoring with a broader workforce sample for a near-practical application and evaluate their effectiveness in real-world scenarios. (Calvetti et al.,2020)

2.3 Research Gap

Previous research has concluded that LEAN Construction is one of the growing fields in managing waste in the construction industry. Many studies focused on managing the waste generated from the inefficient workflow of the workforce on site using LEAN Construction by trying to minimize unnecessary motion, queue, interruption, waiting time of workforce, to maximize the efficiency of the available manpower. However, the process of data collection of such parameters on site remains a tedious, time consuming, and resource exhaustion process. However, according to more recent studies, the field is gearing towards the utilization of real-time monitoring techniques

in the construction field. The implementation of such techniques is getting widely known for monitoring the location and movement of onsite resources. Research has been done in the area of using real-time monitoring techniques in the area of construction safety, followed by fewer research in the area of workers' productivity. Nevertheless, there is a gap in the literature for the implementation of real-time monitoring techniques in collecting data about the workforce in real-time. Also, almost no research has been done regarding the potential implementation and use of spatial statistical analysis and correlation of the collected real-time data. Hence, this primary aim of this research is to develop a framework that fills this literature gap.

CHAPTER 3 – FRAMEWORK DEVELOPMENT FOR FIRST ORDER ANALYSIS

This chapter discusses the development of the framework that is aimed to aid decision makers in utilizing spatial data collected from construction sites. The framework consists of three stages. The first stage is the data collection stage, followed by the data cleaning and preparation stage, and finally the data analysis stage using an algorithmic model. The data collection stage is composed of both manual and semi-automated real-time data collection from a construction site. Data will be collected from personnel on site by using GPS technologies as well as standardized tabulated data records. For the data preparation, a transformation algorithm is used to modify the format of the data so that the data could be processed and analyzed. The data is then analyzed using a programmed algorithm that involves the implementation of several analysis techniques to obtain the frameworks' desired output. The framework composition is shown in Figure 3-1.

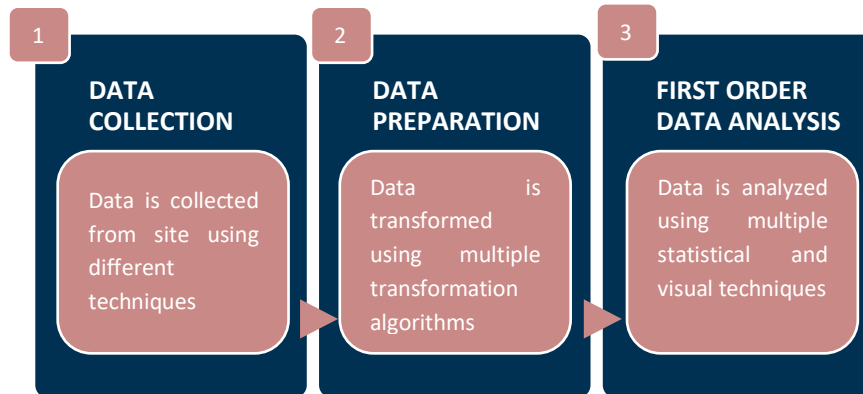


Figure 3-1: Research Framework Stages

3.1 Stage 1 – Data Collection

The first step of collecting the data is to get the data related to the site's geographic location and get the coordinates of the different zones on site, this data shall be referred to as site geographical data (*SGD*). Then, data about the site's performance shall be collected on a periodic basis (daily, weekly, bi-weekly, etc.), these shall be the site periodic data (*SPD*) that be analyzed in the framework. Finally, real-time geospatial data shall be collected from the workforce on site, the dataset shall be known as real-time geospatial data (*RGSD*). All the data is stored on a cloud-based server, such as Microsoft SharePoint (*MSP*), where the second and third stage of data preparation and

analysis take place. The detailed methodology of data collection shall be discussed below.

3.1.1 Site Geographic Data Collection – (SGD)

To collect the site's geographic data, firstly a single predefined coordinate (latitude, longitude) obtained from the projects' tender file. Then, a GIS (Geographic Information System) software or a Geographic browser is used to locate the coordinate. There are multiple software and browsers such as Google Earth Pro or Google Earth Web, NASA's World Wind, ESRI's Explorer for ArcGIS, and GeoFusions's GeoPlayeris. In this framework it is suggested to Google Earth since the browser has a user-friendly interface that could be used by engineers from different fields. The interface is shown in Figure 3-2. In Google Earth, the search box is used to locate the coordinate by simply entering the coordinate in any of the formats below, depending on personal preference or need:

1. Decimal Degrees: such as 37.7°, -122.2°.
2. Degrees, Minutes, Seconds: such as 37°25'19.07"N, 122°05'06.24"W.
3. Degrees, Decimal Minutes: such as 32° 18.385' N 122° 36.875' W.
4. Universal Transverse Mercator: such as 10 S 055974, 4282182.

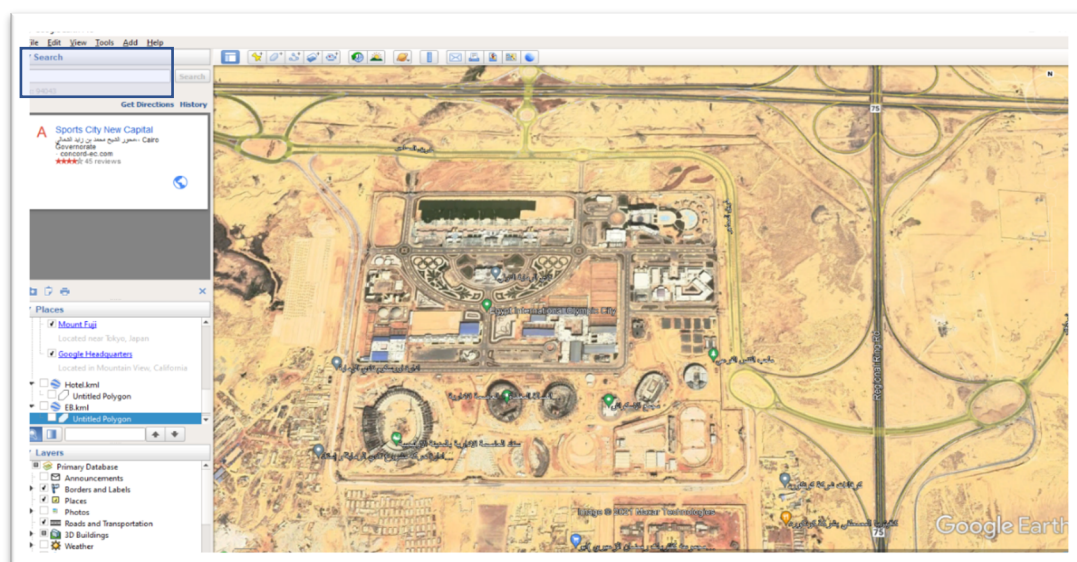


Figure 3-2: Google Earth Interface

It is also recommended to use the World Geodetic Coordinate System (*WGS84*) since it's a universal coordinate system that is commonly used rather than the Egypt 1907/Red belt coordinate system.

After locating the site on Google Earth, the Construction Site Layout Plan (*CSLP*) is used to get the bounding coordinates for the different zones on site as defined in the *CSLP*. Figure 3-3 shows an example of a *CSLP* for a roadway construction site.

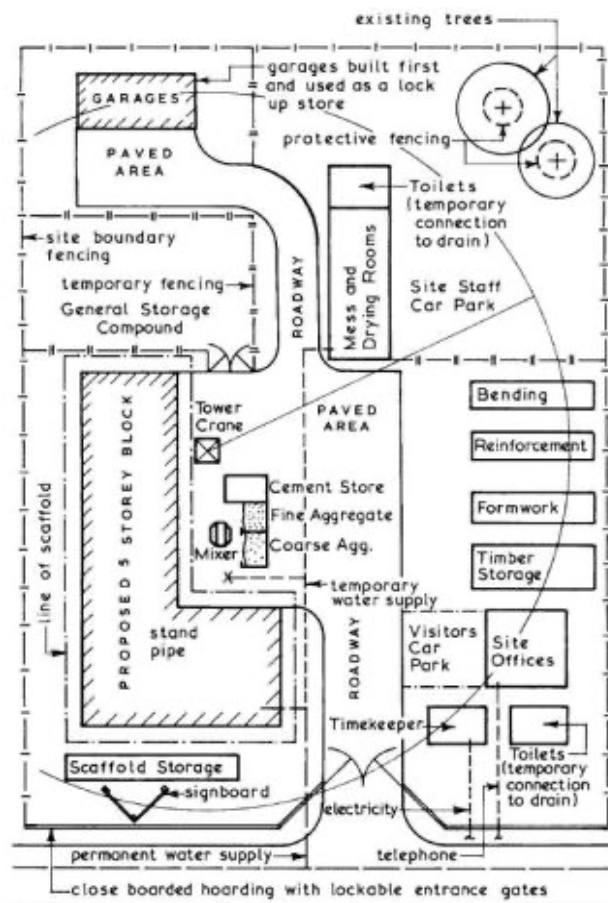
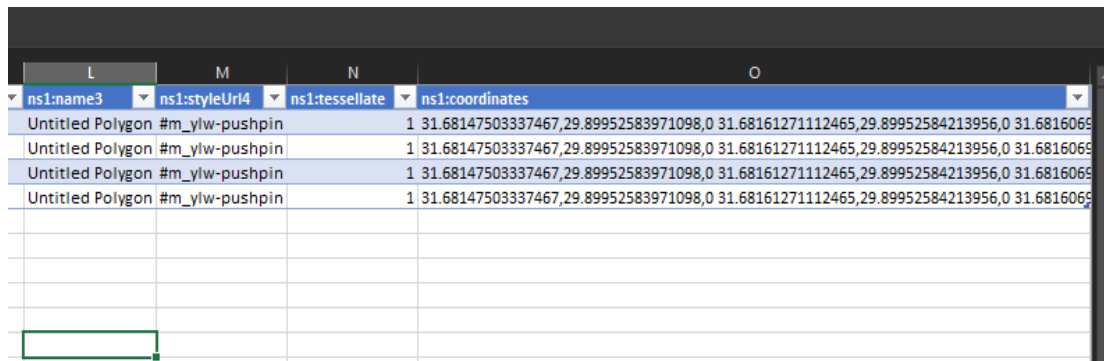


Figure 3-3: Typical Construction Site Layout Plan (The Constructor)

Finally, these coordinates are exported in a xml format as shown in Figure 3-4 and stored on the MSP for the necessary transformation that will be explained in the Data Preparation section.



L	M	N	O
ns1:name3	ns1:styleUrl4	ns1:tessellate	ns1:coordinates
Untitled Polygon #m_ylw-pushpin		1	31.68147503337467,29.89952583971098,0 31.68161271112465,29.89952584213956,0 31.6816069
Untitled Polygon #m_ylw-pushpin		1	31.68147503337467,29.89952583971098,0 31.68161271112465,29.89952584213956,0 31.6816069
Untitled Polygon #m_ylw-pushpin		1	31.68147503337467,29.89952583971098,0 31.68161271112465,29.89952584213956,0 31.6816069
Untitled Polygon #m_ylw-pushpin		1	31.68147503337467,29.89952583971098,0 31.68161271112465,29.89952584213956,0 31.6816069

Figure 3-4: Exported GPS Coordinates from Google Earth

3.1.2 Site Periodic Data Collection – (SPD)

After obtaining the SGD, periodic data regarding the site's safety and quality performance is collected. This method of collection is manual, where both the Health, Safety, and Environment (*HSE*) team and, Quality Control (*QC*) team on site provide the safety and inspection records, respectively, on a periodic basis as explained earlier. Each record shall be referred to the different zones as defined in the Construction Site Layout Plan (*CSLP*). Finally, a tabulated periodic record of the data shall be uploaded by the team members on the same MSP.

3.1.2.1 Safety Records

The HSE team shall periodically document safety accidents, incidents, and near-misses that occurred on site. The documentation provides an indication of the site's safety performance. The data recorded shall include but shall not be limited to:

1. Date: The day, month, and year the incident took place on site.
2. Reference Number: The reference number of the Safety Incident.
3. Time: The time of the day at which the incident took place.
4. Location: The site zone where the incident occurred.
5. Injured Personnel: The name of the person that was injured as a result of the incident.
6. Title: The working title of the person that was injured as a result of the incident.

7. Incident & Injury Description: A brief description of the incident and the resulting injuries if any.
8. Incidents Category: The category of the incidents that took place, whether the incident was falling from height, fire and explosion, Slip and Fall, etc.
9. Damage to Property: The damage to work or property that happened as a result of the incident, if any.
10. Recommended Actions: The action plan suggested to handle the damage done because of the incident and to prevent similar incidents from happening elsewhere on site.
11. Status: Whether the recommended actions have been implemented or not.

The documentation of the data shall be in a tabulated excel sheet in the format shown in Table 3-1.

Table 3-1: Safety Incident Log

Date	No.	Time	Location	Injured Personnel	Title	Incident & Injury Description	Incidents Category	Damage to Property	Recomm. Actions	Status
DD/MM/YY	X _i	HH:MM XM	A	Open/ Closed
DD/MM/YY	X _{i+1}	HH:MM XM	B	Open/ Closed

3.1.2.2 Inspection Records

The QC team shall periodically document both the accepted and rejected inspection requests carried out on site. The documentation provides an indication of the site's quality performance. The higher the number of accepted inspection requests, the better the quality of the work being performed. The data recorded shall include but shall not be limited to:

1. Date: The day, month, and year the inspection took place on site.
2. Reference Number: The reference number of the Inspection Request.
3. Time: The time of the day at which the inspection took place.
4. Location: The construction zone where the inspection was carried out.
5. Construction Element: The Element that was inspected.
6. From: The Site Engineer that submitted the request and has supervised the quality of work executed.

7. To: The QC Engineer that carried out the inspection and determined the quality of the work executed.
8. Drawing Reference: The reference of the drawing used during the inspection process.
9. Code: The code indicates whether the quality of work was acceptable or not. If the Inspection request was given code A, this means the quality of execution was up to standards. Code B indicates an accepted quality of execution with minor defects that shall be repaired. Whereas, Code C, implies unsatisfactory quality of work, thus the element shall be re-executed.

The documentation of the data shall be in a tabulated excel sheet in the format shown in Table 3-2.

Table 3-2: Inspection Requests Log

Date	Ref. No.	Time	Location	Construction Element	From	To	Dwg. Ref.	Code
DD/MM/YY	X _i	HH:MM XM	A	A/B/C
DD/MM/YY	X _{i+1}	HH:MM XM	B	A/B/C

3.1.3 Real-time Geospatial Data Collection – (RGSD)

The last type of data collected from site, is the real-time geospatial data of the site's workforce, both blue and white collar. To collect this geospatial data, workers have to download a GPS tracking application on their work phones and record their daily movement tracks by allowing the application to run as long as they are on site. The workers' then save their tracks and upload them on the MSP for the data to be transformed and analyzed by the model.

3.1.3.1 Smart Phone GPS Tracking Application of Choice & Its Implementation

Multiple GPS Tracking Applications were tested on site during this stage, including but not limited to myTracks – The GPS-Logger, iTrack-GPS Tracking System, etc. The application of choice was myTracks – The GPS-Logger 7.3.0 shown in Figure 3-5. Not only was it recommended by the Iowa State University Geospatial Technology Training

Program, but also it is a free application that can be downloaded on iOS and Android Smartphones thus can be used by most if not all workers on site.

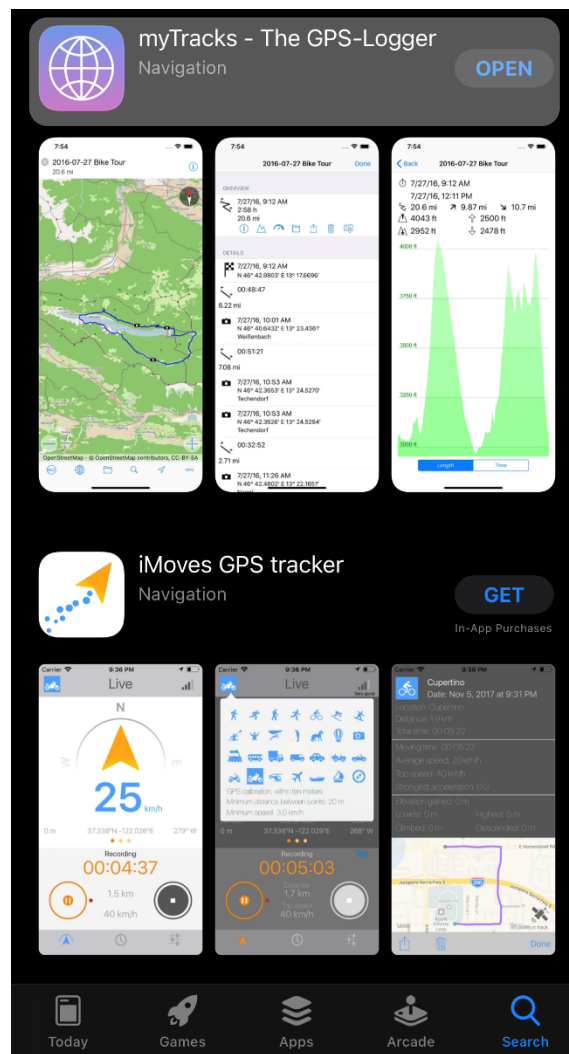


Figure 3-5: Smart Phone GPS Tracking Application (iOS Apple Store)

Other features of myTracks App are:

1. GPS Recording does not require an active internet connection.
2. Tracks are shown on a pixel or vector maps based on OpenStreetMap.
3. Tracks can be organized in folders for different dates.
4. Tracks can be exported as GPX, KML or KMZ files.
5. Special recording mode, called Diary Mode, can be used to create a single track for each day. It uses a power efficient iOS feature called “major location changes”. The diary Mode can be switched on all the time without drilling down the battery.
6. Tracks can be imported to the track library from other applications using the GPX file format.

7. The of GPS Tracking is ± 3 meters.

The application has a user-friendly interface and is easy to navigate. The output from the application can be acquired in a format as shown in Figure 3-6.

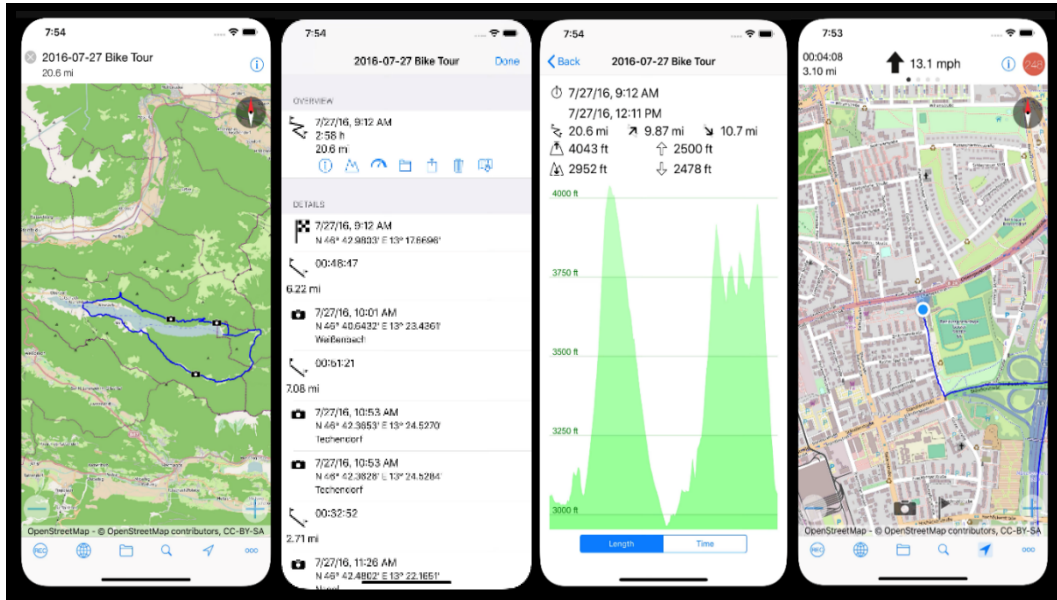


Figure 3-6: Output of using Smart Phone GPS Tracking Application (myTracks iOS Application Interface)

Other than the outputs shown above, the specific coordinates of the recorded tracks can be exported as a gpx file that can be saved to the MSP. A sample of the gpx file is shown in Figure 3-7.

```
E01-(04-10-19).gpx - Notepad
File Edit Format View Help
eed>6.697239511558778</mytracks:speed><mytracks:length>0.001860344308766327</mytracks:length></extensions></trkpt>
181315698275</mytracks:speed><mytracks:length>0.005922623610629786</mytracks:length></extensions></trkpt><trkpt 1
597</mytracks:speed><mytracks:length>0.002266549173510151</mytracks:length></extensions></trkpt><trkpt lat="29.90
acks:speed><mytracks:length>0.001482936760342163</mytracks:length></extensions></trkpt><trkpt lat="29.90070572591
<mytracks:length>0.001954837430364741</mytracks:length></extensions></trkpt><trkpt lat="29.90118584132744" lon="3
th>0.0201569346769382</mytracks:length></extensions></trkpt><trkpt lat="29.90109929817719" lon="31.68391048911184
04373561632</mytracks:length></extensions></trkpt><trkpt lat="29.90098756740792" lon="31.68337907645077"><ele>420
44</mytracks:length></extensions></trkpt><trkpt lat="29.90094331095917" lon="31.68343213389784"><ele>413.04842758
ks:length></extensions></trkpt><trkpt lat="29.90094704090608" lon="31.68338712307781"><ele>413.7796287536621</ele
></extensions></trkpt><trkpt lat="29.90096359516485" lon="31.6834119335112"><ele>418.4141502380371</ele><time>201
sions></trkpt><trkpt lat="29.90099209363563" lon="31.68346490713924"><ele>417.9272727966309</ele><time>2019-04-10
pt><trkpt lat="29.90096845666869" lon="31.68344185690552"><ele>415.4581565856934</ele><time>2019-04-10T06:01:53Z<
lat="29.90092495459123" lon="31.68342643420369"><ele>413.5226707458496</ele><time>2019-04-10T06:02:01Z</time><ext
0093115719957" lon="31.68341453190118"><ele>417.0140037536621</ele><time>2019-04-10T06:02:12Z</time><extensions><
58" lon="31.68340715582639"><ele>416.6854515075684</ele><time>2019-04-10T06:02:20Z</time><extensions><mytracks:sp
31.68343842032522"><ele>420.7391014099121</ele><time>2019-04-10T06:02:25Z</time><extensions><mytracks:speed>7.479
77308649"><ele>419.9107322692871</ele><time>2019-04-10T06:02:30Z</time><extensions><mytracks:speed>6.493574380874
><ele>417.4891624450684</ele><time>2019-04-10T06:02:43Z</time><extensions><mytracks:speed>9.958273818271199</mytr
234153747559</ele><time>2019-04-10T06:02:56Z</time><extensions><mytracks:speed>2.629330727252603</mytracks:speed>
<
Ln 1, Col 1 100% Unix (LF) UTF-8
```

Figure 3-7: gpx File Sample

3.2 Stage 2 – Data Preparation and Cleaning

Since the data has been collected in different formats, some of which are not efficient to process in their original forms, data cleaning and preparation is a necessary step. This step is done automatically using a transformative *splitting and grouping* algorithm with the same programming language that will be discussed later in *Stage 3 – Data Analysis*. Each of the datasets, collected in *Stage 1 – Data Collection*, shall be prepared to produce a certain final output that will be used for the analysis process.

3.2.1 Site Geographical Data Preparation - (SGDP)

Since the SGD is exported in a kml format as explain earlier, the data has to then be modified. By *splitting* the polygon coordinates obtained in the kml file, the separate sets of Easting and Northing coordinates, and the Elevation for each zone are attained. The final output shall be as shown in Table 3-3.

Table 3-3: Site Geographic Data Tabulation Format

Area Code	Ref. No.	Easting	Northing	Elevation
A-XX	X
B-XX	X+1

In the above table, the area code needs to be defined based on the nature of the zone. where the following are the possible areas on site:

- 1. Construction Areas** - Zones where the actual work execution takes place on site.
- 2. Resting Areas** - Areas usually occupied during the break-time of workers or engineers on site. They could be shaded areas or caravans.
- 3. Workshops** - Areas where a fabrication process takes place. On site there are usually rebar workshops, carpentry workshops, mechanical workshops, and electrical workshops.
- 4. Storage Areas** - Zones where material is stored on site.
- 5. Main Caravans** – The caravans where most blue-collar labors are located such as the technical office team, the quantity surveying team, etc.
- 6. Site Boundaries** - These coordinates define the boundary conditions of the site.

For recurrent zones, the code shall be denoted by an increasing number so that each area has a unique code. The development of the area code is shown in Figure 3-8.

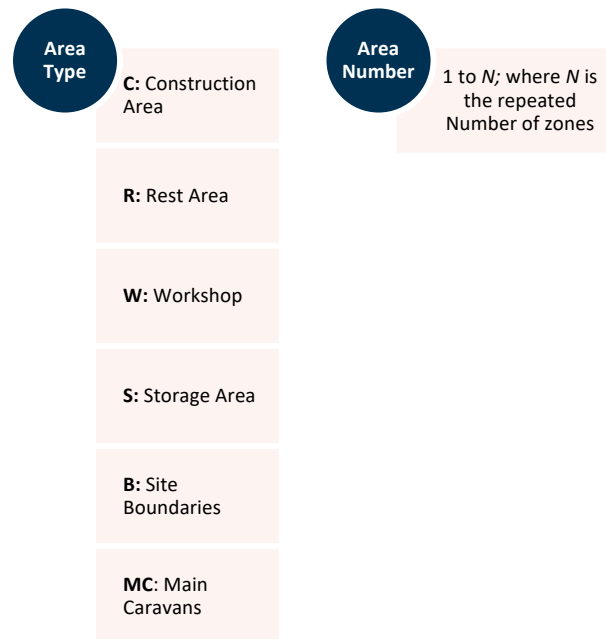


Figure 3-8: Area Unique Code Development

Further classification of the areas is carried out, where each area on site is defined as being either a working area (WA), a resting area (RA) or a travel path (TP). This classification will be necessary as explained later in this chapter. Construction areas, workshops, storage areas, and main caravans are considered as working areas, whereas, as implied by its notation rest areas are considered as resting areas. The remaining undefined site locations within the site boundaries are considered traveling paths. This is done automatically using the area code, according to the area type notation in Figure 3-8, the category of the zone is defined as WA, RA, or TP.

Each Easting and Northing coordinate, in decimal degrees, has a reference number starting for 1 to *n* where *n* is the total number of coordinates on site. Hence, for the first construction area on site, with the first Easting Coordinate 29.912°, the corresponding Northing Coordinate 30.817°, and an elevation of 412 m the tabulated result is as shown in Table 3-4.

Table 3-4: Site Geographic Data Sample

Area_Code	Ref. No.	Easting	Northing	Elevation	Category
C01	1	29.912°	30.817°	412 m	WA

3.2.2 Site Periodic Data Preparation - (SPDP)

For the SPD, the records kept by the site are usually quite extensive and not all data is required for the sake of the analysis in the framework. Therefore, the data is *grouped* in a simpler format as necessary for the analysis.

3.2.2.1 Safety Records

The safety records collected on site are consolidated, for every unique date, to the format shown in Table 3-5, where each incident has been referred to the unique area code where the incident took place.

Table 3-5: Grouped Safety Data Tabulation

Area_Code	No. of Safety Incidents	Date
A-XX
B-XX

The above table constitutes the dataset necessary to evaluate the safety conditions on site and shall be referred to as *Site Safety Data (SSD)*.

3.2.2.2 Inspection Records

Also, the inspection records kept on site are consolidated, for every unique date, to the format shown in Table 3-6, where each record has been referred to the unique area code where the inspection took place.

Table 3-6: Grouped Inspection Data Tabulation

Area_Code	No. of Accepted Inspection Requests	No. of Rejected Inspection Requests	Date
A-XX
B-XX

The above table constitutes the dataset necessary to evaluate the projects' quality of execution and shall be referred to as *Site Quality Data (SQD)*.

3.2.3 Real-time Geospatial Data Preparation - (RGSDP)

Finally, the real-time data collected in the gpx file is then transformed into a more readable and processable format. The gpx files collected from each worker/employee on site are converted to an xml file. The gpx file is *split* at each attribute into different columns, that contain data regarding the employee as well as their coordinates. The data in xml format is shown in Table 3-7.

Table 3-7: Tabulation of Real-time Data After gpx Splitting

Employee ID	Employee Type	Working Activity	Easting	Northing	Elevation	Distance	Speed	Time	Date
A-XX
B-XX

The headers in the table above represent the following:

- 1. Employee ID** - A unique ID given to each employee/worker on site to distinguish employees from one another.
- 2. Employee Type** – Whether the type of work the employee does is supervision or direct construction. Further categorization of the employee type is as follows:
 - Site Engineer – SE
 - Skilled Labor – SL
 - Unskilled Labor – USL
 - Forman – FR

3. Working Activity – The type of activity the employee is involved in, indicating the trade under which the employee is classified. The working activities could be classified as follows:

- Supervision – SP
- Concrete Pouring – CP
- Formwork – FW
- Scaffolding – SC
- Steel Fixing – SF
- Masonry - MS
- Pipe Fitting – PF
- Welding – WL
- Rigging – RG
- Electrical Works – EW
- Etc....

4. Easting and Northing Coordinates: The Easting and Northing coordinates of an employee determine the employee's location on site at a given point in time.

5. Elevation: The Elevation of the employee specifies the height at which the employee is working in meters.

6. Date: The day, month, and year when an Easting and Northing coordinate were recorded.

7. Time: The timestamp refers to the hour, minute, and second at which a specific Easting and Northing coordinate was obtained.

8. Distance: The distance travelled by an employee between time i and time $i+1$ and is measured in meters.

9. Speed: The speed of an employee as the employee moves from one coordinate to another in kilometers per hour.

Hence, for the first Site Engineer on site, on the 12th of February 2020, at 09:10:20 am, with an Easting coordinate of 29.122, and a Northing coordinate of 30.131, the tabulated result is as shown in Table 3-8.

Table 3-8: Sample of Real-time Data Tabulated

Employee ID	Employee Type	Working Activity	Easting	Northing	Elevation	Distance	Speed	Time Stamp	Date
E-01	SE	SP	29.122	30.131	414	0.0012	2.63	09:10:20	11/2/2020

The above table constitutes the dataset containing spatial temporal data of workers the necessary to evaluate the behavior and productivity of workers on site and shall be referred to as *Site Productivity Data (SPD)*.

3.3 Stage 3 – First Order Data Analysis (FODA)

In this stage the transformed data is analyzed using 2-D analysis techniques. These techniques are spatial visual analysis and statistical temporal analysis. The analysis is carried out on the transformed data using an algorithmic model included in appendix A. The model could be developed using any multi-paradigm, object-oriented programming (OOP) language such as Python, C++, Java, etc. In this framework the Python ® Language is used to write the code of the algorithmic model, since it is a high-level programming language that has an English-like syntax that makes it easy to read, learn, and write. The code was developed using the Jupyter™ Notebook, which is an open-source web application that allows the user to create and share documents that contain live codes, equations, visualizations, and narrative text.

3.3.1 Analysis Techniques

The spatial visual analysis and statistical temporal analysis of the data can be carried out using several different mathematical methods. A few of these have been selected to be used as part of the algorithmic model. Firstly, the methods for *Statistical Temporal Analysis* deployed are discussed under this section for the purpose of elaborating on the generic algorithms.

3.3.1.1 Statistical Temporal Analysis

3.3.1.1.1 Mean Calculation

The mean represents the arithmetic average of all observations in a population. This analysis indicates the central value of the observations. The population mean, μ_p , is calculated using Equation (2).

$$\mu_p = \frac{\sum_{i=1}^n p_i}{n} \quad (2)$$

Where n is the total number of observations, and x_i is the value of the observation.

3.3.1.1.2 Median Calculation

Median, $Med(p_o)$, shows the central number of the dataset, by calculating the middle value of the observations in a population. It is obtained using Equation (3).

$$Med(p_o) = \begin{cases} p_o \left[\frac{n}{2} \right] & \text{if } n \text{ is even} \\ \frac{(p_o \left[\frac{n-1}{2} \right] + p_o \left[\frac{n+1}{2} \right])}{2} & \text{if } n \text{ is odd} \end{cases} \quad (3)$$

Where p_o is the ordered list of the values of the observations

3.3.1.1.3 Standard Deviation Calculation

Signifies the dispersion of the value of observations around the mean of the population. The standard deviation, σ_p , is calculated using Equation (4).

$$\sigma_p = \frac{\sqrt{\sum (p_i - \mu_p)^2}}{n - 1} \quad (4)$$

3.3.1.1.4 Moving Average Regression Analysis

Estimates the relationship between a dependent variable and an independent variable. The relationship can be linear or non-linear and can be used to interpolate or extrapolate data. A moving average plot shows the mean of the population, the dependent variable, on one axis, and the independent variable on the perpendicular axis. Hence, the behavior of the mean as the independent variable changes is depicted. Moving averages are unusually used in timeseries analysis, where the pattern of the mean over a period

of time is of interest. In a timeseries, the timeframe is split into equidistant windows over which the moving average is calculated.

Using the same mathematical formulation of Equation (2), the moving average is calculated by determining a value for n over which the mean is calculated progressively. An example of a moving average plot is shown in Figure 3-9.

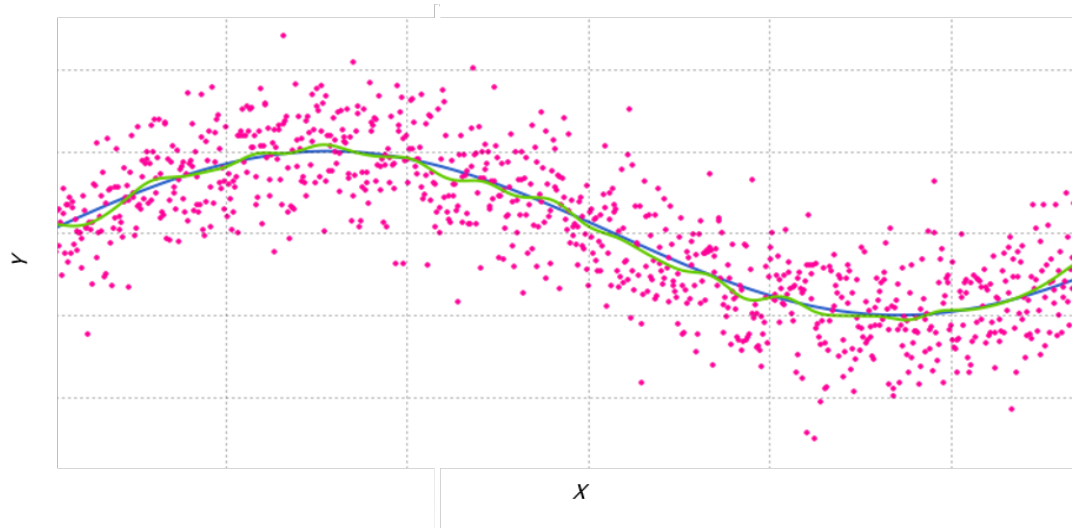


Figure 3-9: Plot of Moving Average

3.3.1.1.5 Point Pattern Analysis

The distribution of spatial observations over a 2-D space is studied using Point Pattern Analysis (PPA). The analysis considers the observations as events that could occur in multiple locations in a space. The occurrence of the datapoints in a few rather than all locations is what is considered a *point pattern*. Point pattern analysis could be done by asking these set of questions:

- *What does the pattern look like?*
- *What is the nature of the distribution of the points?*
- *Is there any structure in the way the locations are arranged over a certain space? Are events clustered or are they dispersed?*

Accordingly, the below measures could be deployed to perform the analysis:

a. Mean Center:

Represents the central tendency of the observations, by estimating the location around which all observations are dispersed. The location of the spatial center is determined using Equation (5). (Yuan et. al., 2020)

$$(\mu_x, \mu_y) = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right) \quad (5)$$

Where, μ_x is the average of all x coordinates, μ_y is the average of all y coordinates, and n is the total number of observed (x, y) coordinates.

b. Median Center:

Signifies the location that minimizes the sum of distances to all observations in a 2-D space. It is an iterative algorithm that starts with any assigned point as the initial median center. The algorithm is shown in Equation (6). (Yuan et. al., 2020)

$$x' = \frac{\sum_i^n \frac{w_i x_i}{d_i}}{\sum_i^n \frac{w_i}{d_i}}, y' = \frac{\sum_i^n \frac{w_i y_i}{d_i}}{\sum_i^n \frac{w_i}{d_i}} \quad (6)$$

Where, x' is the median center of all x coordinates, y' is median center of all y coordinates, w_i is the weight assigned to each chosen median center, and d_i is the distance between a point (x_i, y_i) to the median center from the previous iteration.

This algorithm is repeated until the newly computed median center is not significantly different from the prior one. The mean center and the median center do not usually coincide as shown in Figure 3-10.

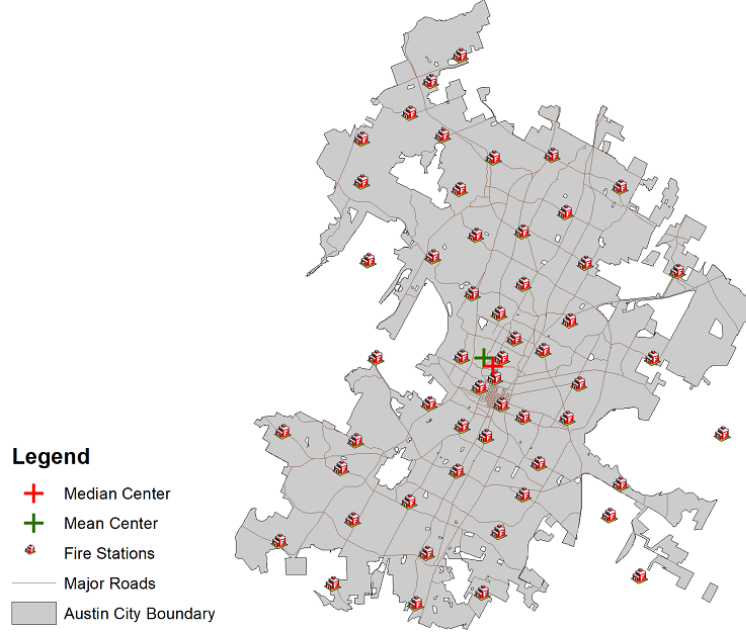


Figure 3-10: Mean and Median Centers of Fire Stations in a City (Yuan et. al., 2020)

c. Standard Distance:

Measures the standard distance which measures the dispersion of the datapoints around the mean center. The standard distance D is calculated using Equation (7). (Yuan et. al., 2020)

$$D = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu_x)^2 + \sum_{i=1}^n (y_i - \mu_y)^2}{n}} \quad (7)$$

d. Standard Deviational Ellipse:

Although the standard distance shows the dispersion of observations, it only calculates an isotropic measure and does not consider directional effect. The standard ellipse in this case indicates the directional dispersion of observations by calculating the standard distances for two perpendicular axes. The center of the ellipse is the mean center, and the major elliptical axis follows the direction of major observation dispersion. The

below equations are used to determine the rotated semi-major (σ_x) and semi-minor (σ_y) axes of a weighted directional distribution. (Yuan et. al., 2020)

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = w \cdot \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad (8)$$

$$\sigma_x = \sqrt{\frac{1}{n} \sum_1^n (\tilde{x}_i \sin \theta + \tilde{y}_i \cos \theta)^2} \quad (9)$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum_1^n (\tilde{y}_i \cos \theta + \tilde{x}_i \sin \theta)^2} \quad (10)$$

Where W is the weight matrix and the rotation angle θ is calculated using Equation (11).

$$\tan \theta = \frac{(\sum_{i=1}^n \tilde{x}_i^2 - \sum_{i=1}^n \tilde{y}_i^2) + \sqrt{(\sum_{i=1}^n \tilde{x}_i^2 - \sum_{i=1}^n \tilde{y}_i^2)^2 + 4 \sum_{i=1}^n \tilde{x}_i \tilde{y}_i}}{2 \sum_{i=1}^n \tilde{x}_i \tilde{y}_i} \quad (11)$$

The standard deviational ellipse should look similar to the one shown in Figure 3-11.

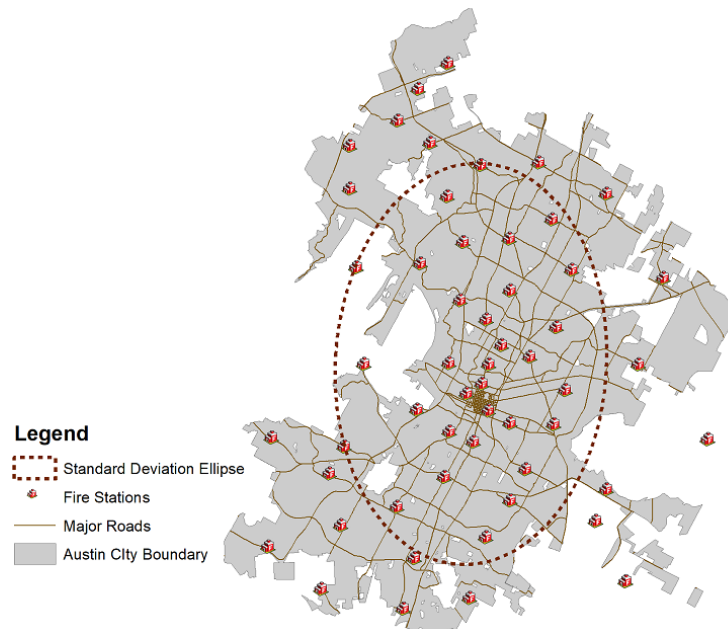


Figure 3-11: Standard Deviation Ellipse of Fire Stations in a City (Yuan et. al., 2020)

e. Nearest Neighbor Distance:

Another important type of PPA is spatial randomness and clustering. The analysis indicates whether the point pattern is completely random or if there are clusters of observations centered around different mean centers. One of the techniques used to test spatial randomness and clustering is the Nearest Neighbor Distance – (NND). NND is the distance between a point and the point closest to it. The mean of the NND calculated between all point pairs in a dataset is used as the global indicator to measure the point pattern of the dataset. The mean of the NND can be compared with the NND expected from points following complete spatial randomness.

The mean of NND of a dataset that follows complete spatial randomness from a Poisson distribution can be seen in Figure 3-12.

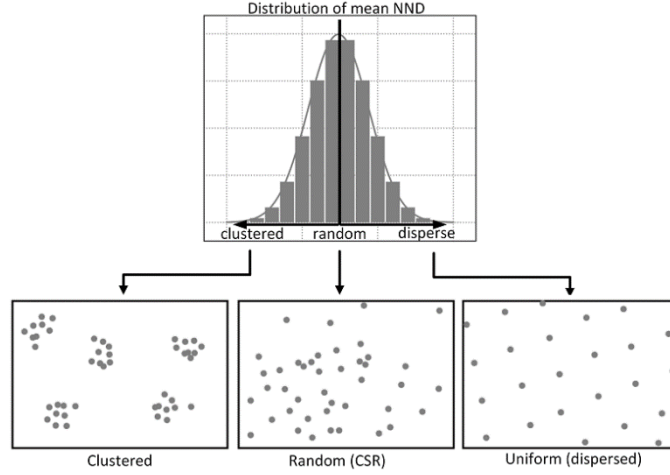


Figure 3-12: Poisson Distribution for Complete Spatial Randomness (Yuan et. al., 2020)

The probability that the dataset does not exhibit complete spatial randomness is indicated by the standard deviation or the z-score of the point pattern. In order to test the point pattern of the population, firstly, the mean NND (μ_{NND}) of the population is calculated using Equation (12). (Yuan et. al., 2020)

$$\mu_{NND} = \frac{\sum_{i=1}^n d_i}{n} \quad (12)$$

Where d_i is the NND of point with its nearest neighboring point.

Then the expected mean NND of a population in complete spatial randomness is calculated as:

$$\mu_{CSR} = \frac{0.5}{\sqrt{\frac{n}{A}}} \quad (13)$$

Where A is the area of the minimum bounding box of the point set. The ratio between μ_{NND} and μ_{CSR} indicates the distribution of data, if the ratio is < 1 then the data is dispersed, if the ratio is > 1 the data is clustered. Finally, the z-score of the mean NND is calculated as:

$$z = \frac{\mu_{NND} - \mu_{CSR}}{DE} \quad (14)$$

Where DE is obtained using Equation (15). (Yuan et. al., 2020)

$$DE = \frac{0.261356}{\sqrt{\frac{n^2}{A}}} \quad (15)$$

The z-score then indicates the level of confidence of the detected point pattern. The higher the z-score the more significant the pattern is.

Although the mean NND can be used as a measure of clustering, it provides limited data about the complexity of the point pattern at multiple spatial scales. Hence, the use of a *G-distance function* gives more information about the degree of clustering of observations at different distances d . G is defined as:

$$G(d) = \frac{\sum(D_{xy} < d)}{n} \quad (16)$$

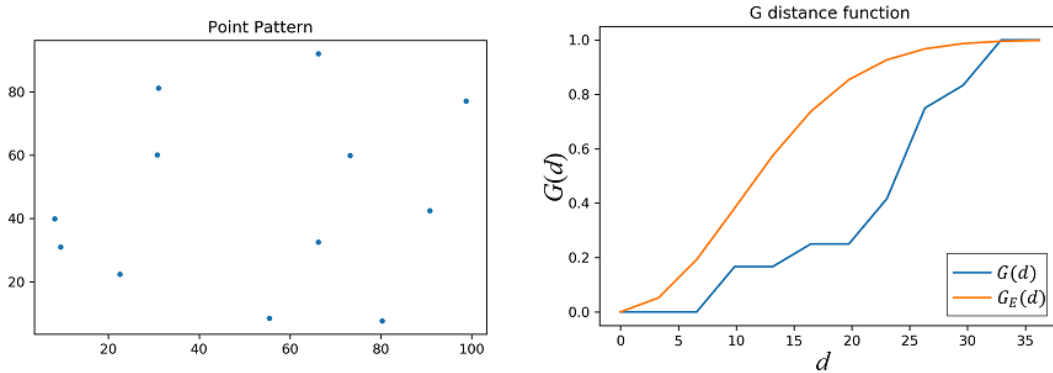


Figure 3-13: G-function Plot (Yuan et. al., 2020)

Figure 3-13 shows the *G-distance* plot of a point pattern, where $Gg(d)$ is the G envelope of a point pattern that is expected to follow complete spatial randomness, and $G(d)$ is for the point pattern. The shape of the G function represents how a point pattern clusters.

If the observations are clustered, the G function increases rapidly at shorter distances, and vice versa.

f. Clustering Analysis:

If the point pattern has been identified to of a clustered type, clustering analysis is used in attempts define clusters found within a population using clustering algorithms. A list of the 10 most common clustering algorithms is as follows:

- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Mixture of Gaussians

Each algorithm has a different approach to define clusters within a population. In this framework the BIRCH algorithm is deployed as it is efficient when it comes to handling larger datasets as is the case here. Also, BIRCH handles ‘noise’ observations effectively. BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies incrementally and dynamically clusters observations to try and produce the best quality clustering. It is an unsupervised data mining algorithm that requires a single scan of the population. The algorithm uses the n number of observations, and the chosen number of clusters K and begins by building a clustering feature (CF) tree out of the given data points. The CF is calculated using Equation (17). (Ramadhani et. al., 2020)

$$CF = (n, \overrightarrow{LS}, SS) \quad (17)$$

Where, \overrightarrow{LS} is the linear sum of the attribute value x and is calculated as follows:

$$\overrightarrow{LS} = \sum_{i=1}^n \vec{x}_i \quad (18)$$

SS is the squared sum of datapoints and is calculated as follows:

$$SS = \sum_{i=1}^n \vec{x}_i^2 \quad (19)$$

Using the CF , the CF-tree, which is a height balanced tree with a branching factor B and threshold distance T , is constructed as shown in Figure 3-14. Each non-leaf node has at most B non-leaf entries in the form of $[CF_i, child_i]$. $Child_i$ refers to the i -th child node. A non-leaf node shows a cluster constituted of all sub-clusters represented by its entries. A leaf node has entries in the form of $[CF_i]$, and the leaf nodes are linked together by previous and next pointers. All entries to the leaf node meet the following condition:

$$r = \sqrt{\frac{SS}{n} - \left(\frac{\overline{LS}}{n}\right)^2} < T \quad (20)$$

Where, r is the radius of the cluster.

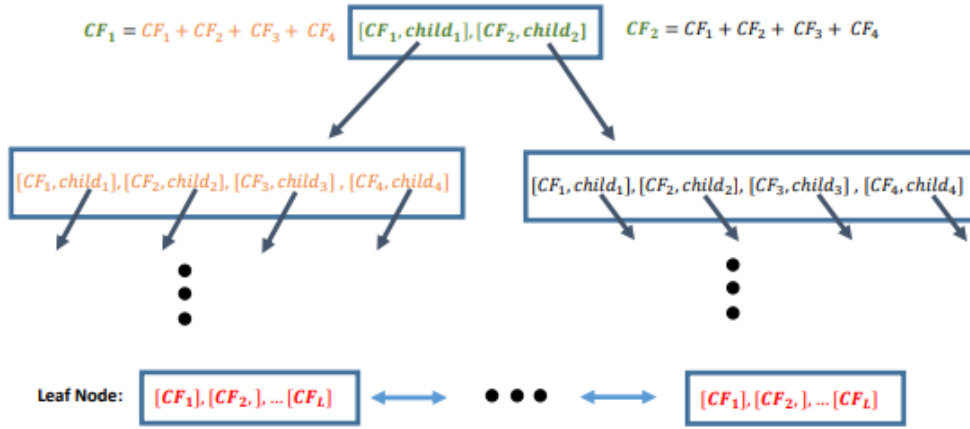


Figure 3-14: BIRCH Clustering CF-Tree Construction

After the clustering analysis has been finalized and the clusters have been identified, the clusters are visualized by representing each cluster by a different color as shown in Figure 3-15.

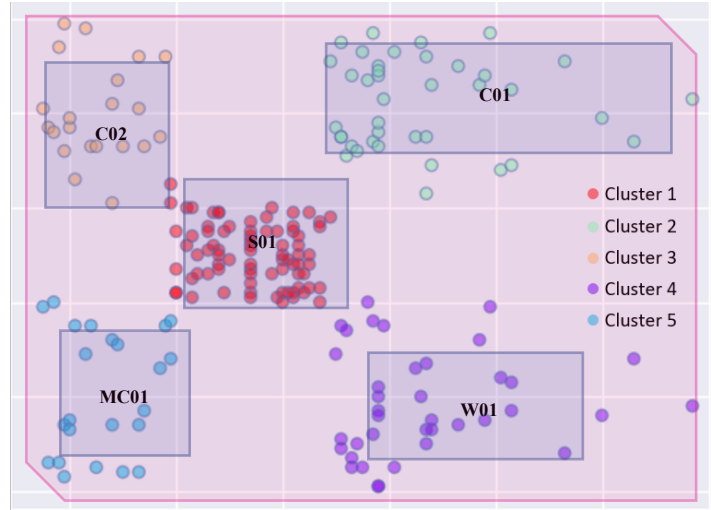


Figure 3-15: Birch Clustering Sample Output (Thecleverprogrammer)

g. Quadrat Density:

In quadrat density analysis, the 2-D space is divided into sub-regions and the point density is calculated for each sub-region. The density of observations within a quadrat is simply calculated by counting the number of observations within this quadrat. The quadrat density is shown in Figure 3-16, where values within each quadrat represent the point count within the quadrat.

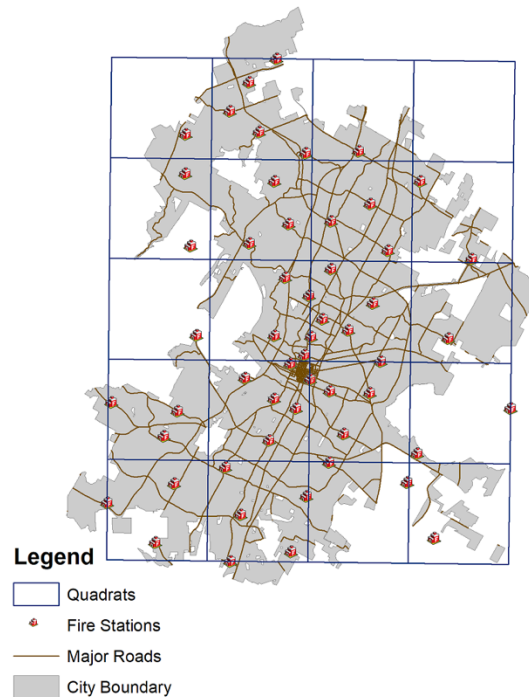


Figure 3-16: Quadrat Density of Fire Stations in a City (Yuan et. al., 2020)

The spatial randomness and clustering of the points can also be examined by using quadrat density. A chi-squared test, χ^2 , is used to determine whether the point pattern of the observations is random or clustered by referring to the densities within the quadrats. The χ^2 test is a statistical hypothesis test is used to determine whether this is a statically major difference between the expected densities and the observed densities. The standard application of the test includes classifying the point densities into mutually exclusive events. If the null hypothesis, that there are no deviations between the observed densities, then the test statistic follows a χ^2 frequency distribution. Assuming that the null hypothesis is true, the test is used to evaluate how closely related the observed densities are. Meaning that the observed densities follow a spatial random point pattern. To determine whether the null hypothesis is true, the significance value, p -value, of a normally distributed set of datapoints, is obtained. The p -value is the probability of variation between the observed densities and the expected densities as per the null hypothesis. Thus, a p -value greater than 0.05 means that the null hypothesis is accepted, i.e., the densities of the workers follow the behavior of complete spatial randomness, and vice versa. The p -value is obtained from a t -distribution. Using the relative t , calculated using Equation (21). (Yuan et. al., 2020)

$$t = \frac{\frac{\sum d}{n}}{\sqrt{\frac{\sum d^2 - (\frac{\sum d}{n})^2}{(n-1)(n)}}} \quad (21)$$

where d is calculated as follows,

$$d = \sqrt{(x_i - x_{i+n})^2 + (y_i - y_{i+n})^2} \quad (22)$$

the p-value can be obtained from the normal distribution graph in Figure 3-17.

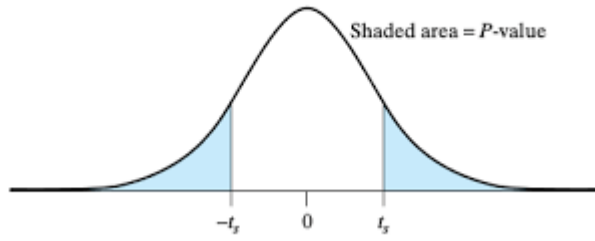


Figure 3-17: Normal Distribution Graph

A major downfall of this technique is that the shape and size of the chosen sub-regions highly affects the p -value. This is especially true for a space that does not have rectangular boundaries as the expected densities will be dissimilar to those of a rectangular boundary.

h. Kernel Density:

Another measure of the observation density is the kernel density. It estimates the local density of observations in a continuous manner by counting the frequency of observations within a region. A Kernel Density Estimate (KDE) calculated using Equation (23), determines the densities of the occurrence of observations. (Chen, 2017)

$$q_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{p_i - p}{h}\right) \quad (23)$$

Where $K(x)$ is the Kernel Function, and h is the bandwidth. The Kernel Function is usually a smooth, symmetric function, where each observation x_i is smoothed using a density bump, then the sum of all density bumps is used to obtain the final density estimate as shown in Figure 3-18.

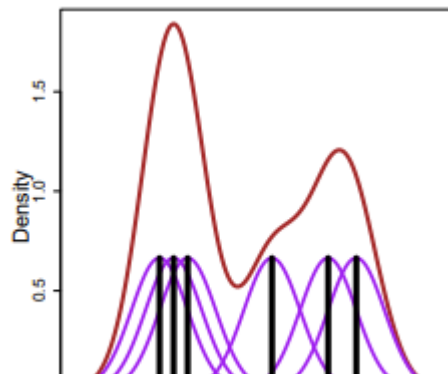


Figure 3-18: KDE Bump Smoothing (Chen, 2017)

In the figure above, 6 datapoints are smoothed, where smooths each data point is smoothed into smaller purple density bumps and then the sum of these bumps is used to obtain the final density estimate represented by the brown density curve.

The choice of bandwidth h affects the smoothness of the Kernel Function. Different bandwidths produce different smoothed densities as shown in Figure 3-19.

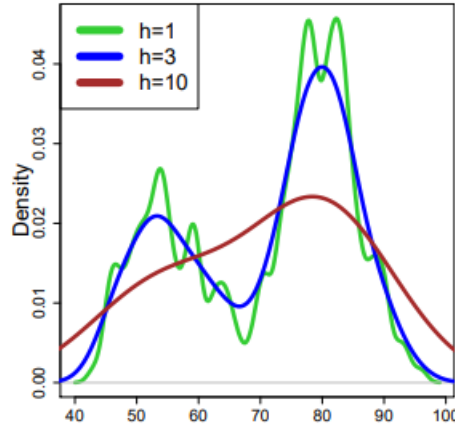


Figure 3-19: Bandwidth Effect on KDE (Chen, 2017)

From the above figure it is depicted that choosing an h that is too small, yields an unsmoothed curve – the green curve, this is known as undersmoothing. On the other hand, when h is too large the curve is too smooth – brown curve, this is called over-smoothing.

Another factor that affects the output of the KDE is the Kernel Function $K(p)$ itself. Generally, there are 3 features to a Kernel Function, these are:

1. $K(p)$ is symmetric.
2. $\int K(p) dp = 1$.
3. $\lim_{x \rightarrow -\infty} K(p) = \lim_{x \rightarrow +\infty} K(p) = 0$

In particular, the second feature is needed to guarantee that the KDE is a probability density function. There are several types of Kernel Function that are commonly used: uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic, and cosine. The Kernel Function deployed in this framework is of type Gaussian, calculated using Equation (24). (Chen, 2017)

$$K(p) = \frac{1}{\sqrt{2\pi}} e^{\frac{-p^2}{2}} \quad (24)$$

The final output of the KDE application is a probability distribution density graph as shown in Figure 3-20. The y-axis of the plot shows the probability density of, and the x-axis represents the values of x .

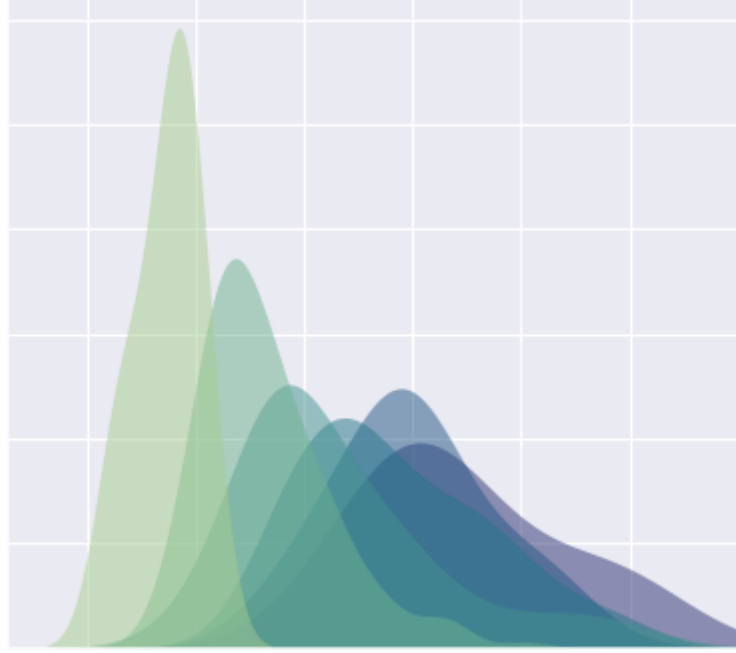


Figure 3-20: KDE Final Plot (Waskom, 2021)

KDE is not only used for analyzing the density of a single variable. It can be used to get the density of 2 or more variables. In this framework, bivariate analysis is of interest. Following the same mathematical formulas of KDE for variable x and variable y , a bivariate contour plot can be obtained as seen in Figure 3-21.

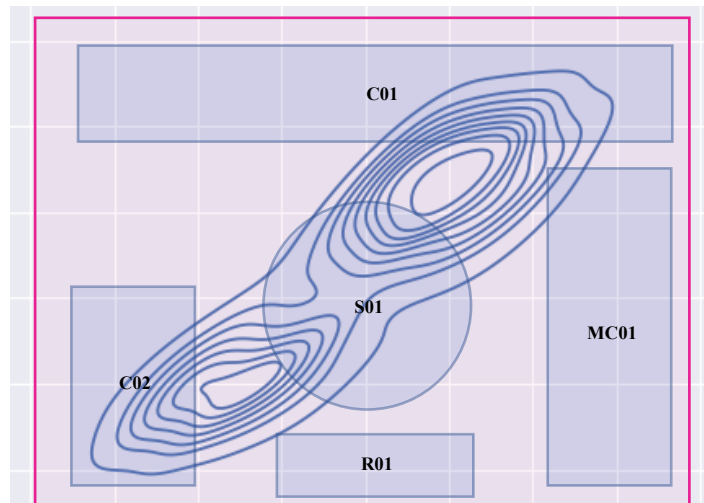


Figure 3-21: KDE Bivariate Contour Plot (Waskom, 2021)

In the KDE contour plot, the axes represent the variable and the lines drawn are known as contour lines. Contour lines give an indication of the number of parts that the probability density function has been split into. Hence, four contour lines mean that the density function has been split into five parts. Each part represents the density outside the contour lines. Thus, each contour line is defined to be at a certain density from $[0,1]$. For example, a contour line set at 0.2 means that 20% of the probability density lies outside the 20% contour line. This means that the smaller the area between the contour lines at a given location, the higher the probability densities of observations in that location. KDE plots can then be transformed into heatmaps by providing a color scheme to the contoured area, yielding the output shown in Figure 3-22.

3.3.1.2 Statistical Visual Analysis

3.3.1.2.1 Heatmaps

Visual/graphical representation of datapoints where values are depicted by colors. The warm-cool color gradient in heatmaps is used to show the behavior of the datapoints. A heatmap is simply a color contoured KDE plot. The color contour goes from colors of highest intensity representing the largest densities, to the lowest intensity representing the smallest densities. A sample of a 2-D contour plot is shown in Figure 3-22.

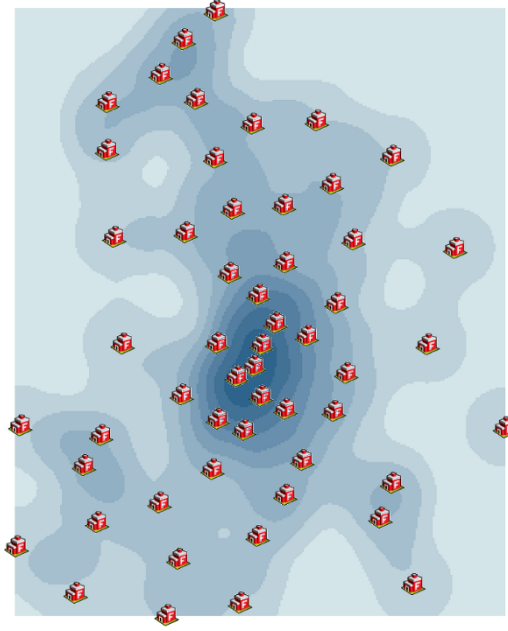


Figure 3-22: Heatmap of Fire Stations in a City (Yuan et. al., 2020)

The contour plot is then referred to as the *Heatmap*.

3.3.1.2.2 Voronoi Diagrams

Voronoi is a partitioning technique that is commonly used with spatial data. The technique shows the closeness of each datapoint to the datapoint next to it in a plane. The plane is divided into a set of Voronoi regions in the shape of polygons. Each polygon encompasses the set of observations closest to the observation around which the polygon was constructed. Hence, a Voronoi diagram is a visualization tool that is a data-driven tessellation of a plane. Tessellation or tiling is the processing of covering a plane, in this case a spatial plane, using one or more geometric shapes. The most common type of tessellation is tiling using polygons as seen in Figure 3-23.

The points that generate the Voronoi diagram shown above are called “generators”. Each generator creates its own polygon known as a Voronoi “cell”. The cell embodies the space in the plane that is closest to its generator. The shared boundaries of cells signify the space that is equidistant to the generators. The cell of a generator R_k is defined for each cell using Equation (25). (Ferrero, 2011)

$$R_k = \{g \in \mathbb{R}^2 \mid |g - m_i| \leq |g - m_i| \text{ for all } m \in M\} \quad (25)$$

Where M is the set of (x, y) coordinates $\{m_l = (x_l, y_l), \dots, p_n = (x_n, y_n)\}$ in \mathbb{R}^2 , and g is the generator.

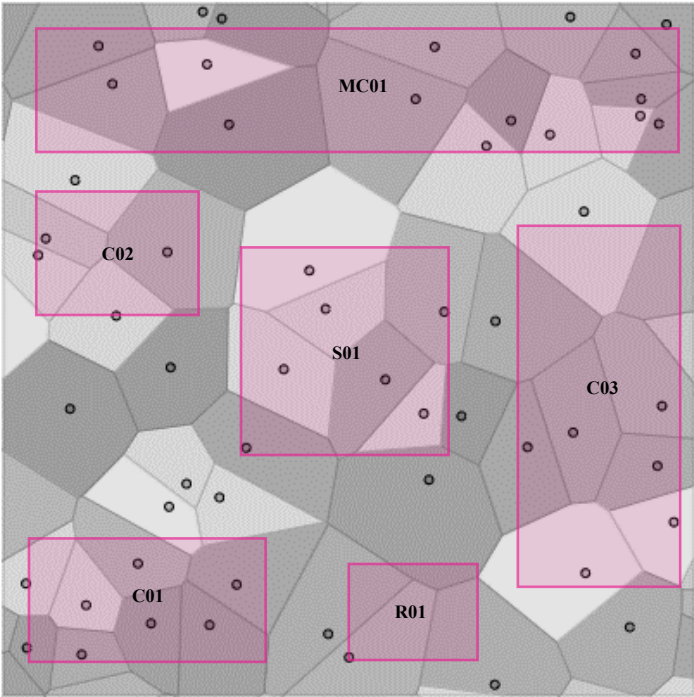


Figure 3-23: Voronoi Diagramme (Elm Packages)

All of the discussed analysis methods can be used to analyze the different data sets collected. Not all the methods can be or should be applied to all data. Accordingly, the implementation of the methods in the framework are shown in Table 3-9.

Table 3-9: Analysis Implementation Matrix

	VISUAL		STATISTICAL				
	Heatmaps	Voronoi Diagram	Point Pattern Analysis	Regression Analysis	Mean	Median	Standard Deviation
Site Periodic Data	●			●	●	●	●
Real-time Data	●	●	●				

3.3.2 Analysis Techniques Implementation and Outputs

The above-mentioned analyses methods are then applied to the transformed data. The application of the methods yields multiple outputs for each dataset. The datasets are categorized by type as mentioned in Stage 2 – Data Preparation and Cleaning, and the different methods are applied to each dataset according to the Implementation matrix in Table 3-9. Hence, the analysis is categorized according to the dataset undergoing the analysis. These categories are (1) Safety Data Analysis (*SDA*), (2) Quality Data Analysis (*QDA*), and (3) Productivity Data Analysis (*PDA*). This sub-section discusses the detailed implementation of the methods as well as the possible outputs that could be obtained under each of the categories.

3.3.2.1 Safety Data Analysis (SDA)

By means of the data collected and the analysis methods applied as mentioned earlier, the possible outputs of analyzing safety incidents on a construction site are:

1. Site Safety Performance (SSP) which is provided by:
 - a. *Temporal Mean.*
 - b. *Temporal Median.*
 - c. *Temporal Standard Deviation.*
2. Site Safety Behavior (SSB) which is provided by:
 - a. *Moving Average Linear Regression.*
3. Site Safety Risk Zones (SSRZ) using heatmaps that show:
 - a. *Safety High-risk Zones*
 - b. *Safety Medium-risk Zones*
 - c. *Safety Low-Risk Zones*

3.3.2.1.1 Site Safety Performance (SSP)

Firstly, below is a list of statistical methods used to analyze the temporal data collected about the safety incidents on site. This analysis indicates the site's performance in terms of safety. The data is collected and analyzed over a time period t . t could be measured in days, weeks, months, and years.

a. Mean Calculation:

Gives the average number of safety incidents over a period of time. The average, μ_S , is calculated using Equation (2), replacing p with S , the number of safety incidents on site. Where n represents t which is the total number of days over which the incidents were recorded. A construction site with a lower mean of safety incidents reflects better safety performance on site.

b. Median Calculation:

Gives the middle datapoint of safety incidents over a period of time. The median, $Med(S_o)$ is calculated using Equation (3), where p_o becomes S_o and is the ordered list of the number of safety incidents. The less the median becomes, the better the site is performing in terms of safety.

c. Standard Deviation Calculation:

Indicates the distribution of datapoints around the central mean of the recorded safety incidents over a certain period of time. The standard deviation, σ_S , is calculated using Equation (4). The value of the standard deviation signifies the stability of the safety conditions on site. A low value of the standard deviation signifies consistency of the safety conditions on site and that the conditions are generally maintained at a certain standard. The standard is determined from the mean as explained earlier.

3.3.2.1.2 Site Safety Behavior (SSB)

a. Moving Average Linear Regression:

This determines the relationship between the time elapsed in a project and the number of safety incidents that occurred on site. This relationship could then be used to extrapolate the expected number of safety incidents that might occur during the remaining the project life. The regression line is drawn based on a moving average algorithm, calculated using Equation (2) using t to represent the window of a working week on site. An example of a moving average linear regression is shown in Figure 3-9.

3.3.2.1.3 Site Safety Risk Zones (SSRZ)

The second type of analysis technique used the visual analysis, where density heatmaps are used to represent the safety risk zones on site. Heatmaps are generated using Kernel Density Estimation explained under 3.3.1.1.5 (h) – Kernel Density.

b. Kernel Density Estimation – KDE:

The Kernel Density Estimate (KDE) calculated using Equation (23), calculates the densities of the occurrence of safety incidents on site. Where $K(S)$ is the Kernel Function of the safety incidents, n is represented using S , and h is the bandwidth. The Kernel Function used for safety incidents is Gaussian and is calculated using Equation (24). The densities calculated are then plotted against time t .

Finally, a 2-D color contour plot is then generated by adopting the density function produced by the KDE. The produced heatmap is shown in Figure 3-24.

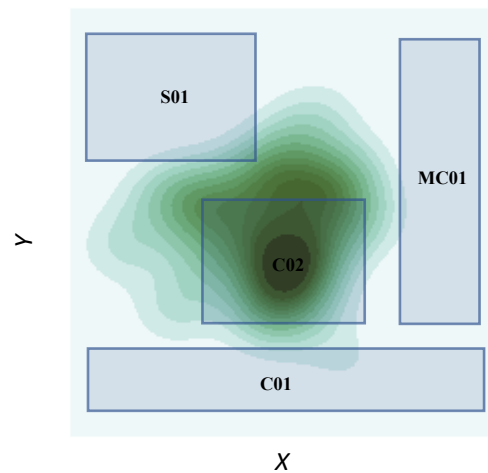


Figure 3-24: Heatmap Sample (Waskom, 2021)

In the above plot, the X-axis represents the Easting co-ordinates, the Y-axis represents the Northing co-ordinates, and the contouring shows the densities of the safety incidents that occurred on site either at a given t_i or over a period of time t as specified earlier.

Using $\mu = \mu_s$ and $\sigma = \sigma_s$, we can identify the following on a construction site:

- a. Safety High-Risk Zones – where the high-risk zone density (*HRZD*) is defined by densities falling within the range of $\mu_s + 2 \sigma_s$.
- b. Safety Medium-Risk Zones – where the medium-risk zone density (*MRZD*) is defined by densities falling within the range of $\mu_s + \sigma_s$ and $\mu_s - \sigma_s$.
- c. Safety Low-Risk Zones – where the low -risk zone density (*LRZD*) is defined by densities falling within the range of $\mu_s - 2 \sigma_s$.

3.3.2.2 Quality Data Analysis (QDA)

By means of the data collected and the analysis methods applied as mentioned earlier, the possible outputs of analyzing inspection requests on a construction site are:

1. Site Quality Performance (SQP) which is provided by:
 - a. *Temporal Mean.*
 - b. *Temporal Median.*
 - c. *Temporal Standard Deviation.*
4. Site Quality Behavior (SQB) which is provided by:
 - a. *Moving Average Linear Regression.*
5. Site Quality Zones (SQZ) using heatmaps that show:
 - a. *High-quality Zones*
 - b. *Medium-quality Zones*
 - c. *Low-quality Zones*

3.3.2.2.1 Site Quality Performance (SQP)

Firstly, below is a list of statistical methods used to analyze the temporal data collected about the accepted and rejected inspection requests on site. This analysis indicates the site's performance in terms of quality of work executed. The data is collected and analyzed over a time period t . t could be measured in days, weeks, months, and years.

a. Mean Calculation:

Gives the average number of accepted and rejected inspection requests over a period of time. The average, μ_R , is calculated using Equation (2), replacing p with R , the number of accepted/rejected inspection requests. Where n represents t which is the total number of days over which the incidents were recorded. A construction site with a higher of accepted requests reflects better quality of work than that on a site with a lower mean.

b. Median Calculation:

Gives the middle datapoint of inspection requests over a period of time. The median, $Med(R_o)$ is calculated using Equation (3), W =where R_o is the ordered list of the number of accepted and rejected inspection requests recorded. The less the median becomes, the better the site is performing in terms of quality.

c. Standard Deviation Calculation:

Indicates the distribution of datapoints around the central mean of the recorded number of accepted and rejected inspection requests over a certain period of time. The standard deviation, σ_R , is calculated using Equation (4). The value of the standard deviation signifies the consistency of the quality of work executed on site. A low value of the standard deviation signifies reliability of the quality of work executed and that the quality is generally maintained at a certain standard. The standard is determined from the mean as explained earlier.

3.3.2.2.2 Site Quality Behavior (SQB)

b. Moving Average Linear Regression:

This determines the relationship between the time elapsed in a project and the number of accepted and rejected inspection requests on site. This relationship could then be used to extrapolate the expected rate of acceptance or rejected of a request that might be submitted during the remaining the project life. The regression line is drawn based on a moving average algorithm, calculated using Equation (2) using t to represent the window of a working week on site. An example of a moving average linear regression is shown in Figure 3-9.

3.3.2.2.3 Site Quality Zones (SQZ)

The second type of analysis technique used the visual analysis, where density heatmaps are used to represent the quality zones on site. Heatmaps are generated using Kernel Density Estimation explained under 3.3.1.1.5 (h) – Kernel Density.

c. Kernel Density Estimation – KDE:

The Kernel Density Estimate (KDE) calculated using Equation (23), calculates the densities of the occurrence of accepted and rejected inspection requests on site.

Where $K(R)$ is the Kernel Function, n is replaced with R , the number of accepted and rejected inspection requests that were submitted, and h is the bandwidth. The Kernel Function used for accepted and rejected inspection requests, is Gaussian and is calculated using Equation (24).

Finally, a 2-D color contour plot is then generated by adopting the density function produced by the KDE. The produced heatmap is shown in Figure 3-25.

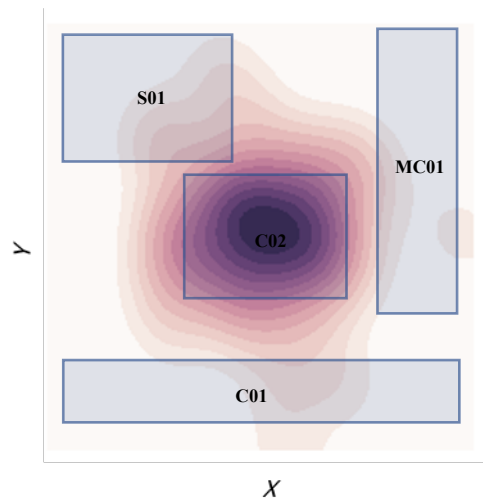


Figure 3-25: Heatmap Sample (Waskom, 2021)

In the above plot, the X-axis represents the Easting co-ordinates, the Y-axis represents the Northing co-ordinates, and the contouring shows the densities of the accepted and rejected inspection requests that occurred on site either at a given t_i or over a period of time t as specified earlier.

Using $\mu = \mu_R$ and $\sigma = \sigma_R$, we can identify the following on a construction site:

- d. High-quality Zones – where the high-quality zone density (*HQZD*) is defined by densities falling within the range of $\mu_R + 2 \sigma_R$.
- e. Medium-quality Zones – where the medium-quality zone density (*MQZD*) is defined by densities falling within the range of $\mu_R + \sigma_R$ and $\mu_R - \sigma_R$.
- f. Low-quality Zones – where the low-quality zone density (*LQZD*) is defined by densities falling within the range of $\mu_R - 2 \sigma_R$.

3.3.2.3 Productivity Data Analysis (PDA)

Given the real-time GPS transformed coordinates of the workers', multiple analyses could be performed on such data. The results from the analyses could give decision makers valuable information about the workers' spatial behavior on site, and how such behavior is affecting the project's performance. Then, upon collection of the required real-time data of workers' temporal spatial coordinates, the analysis yields the following outputs:

- 1. Central Tendency of Workers on Site
- 2. Spatial Randomness and Clustering of Workers on Site
- 3. Workers' Density on Site using Heatmaps that determine:
 - a. High-activity Zones
 - b. Medium-activity Zones
 - c. Low-activity Zones
- 4. Workers' Time Distribution

3.3.2.3.1 Central Tendency of Workers on Site

Point pattern analysis (*PPA*) is used to analyze the spatial temporal data of workers collected from site and indicate the central tendency or dispersion of workers on site. The location of a worker is considered an event, the worker could be located anywhere on site. However, given that the worker is located in a specific place at a specific point in time is the point pattern of the worker. The location of the worker also indicates the productivity of the worker as will be explained later in this section.

The point pattern is significant in the field of construction as it provides major information about the behavior of personnel on site based on their locations. Their behavior is modeled against other site performance measures to detect whether the aforementioned behavior has an effect on the projects' performance or not. After determining the existence of a relationship between the workers conduct and the project's performance, it could then be identified either as a positive or negative relationship.

Hence, for a set of random workers spatial temporal datapoints the following PPA could be carried out by calculating the following measures:

a. Mean Center:

In PPA the mean center of a dataset of workers spatial temporal coordinates gives an indication about the central tendency and dispersion of the workers on site. It provides an estimation of the location around which all points are scattered. The estimate of this center is provided by averaging the Easting and Northing Coordinates of all points in the dataset. The mean center could be calculated for daily, weekly, bi-weekly, monthly, quarterly, or yearly data. In any case, the datapoints are grouped in the required analysis timeframe and the mean center is calculated using the Equation (26).

$$(\mu_j, \mu_k) = \left(\frac{\sum_{i=1}^n j_i}{n}, \frac{\sum_{i=1}^n k_i}{n} \right) \quad (26)$$

Where, j is the Easting coordinate, k is the Northing coordinate, μ_j is the average of all Easting coordinates, μ_k is the average of all Northing coordinates, and n is the total number of coordinates. This signifies the central density of workers on site at a given point in time.

b. Median Center:

Another measure of the data's central tendency is the median center. This is the location to which the sum of all distances is minimized. The algorithm is shown in Equation (27).

$$j' = \frac{\sum_i^n \frac{w_i j_i}{d_i}}{\sum_i^n \frac{w_i}{d_i}}, k' = \frac{\sum_i^n \frac{w_i k_i}{d_i}}{\sum_i^n \frac{w_i}{d_i}} \quad (27)$$

Where, j' is the median center all Easting coordinates, k' is median center of all Northing coordinates, w_i is the weight assigned to each chosen median center, and d_i is the distance between a point (j_i, k_i) to the median center from the previous iteration.

The median center is considered a better robust indicator of the central tendencies of workers as it is less impacted by outliers than the mean center is. This signifies the weighted central density of workers on site at a given point in time.

c. Standard Distance:

The standard distance provides a measure of how dispersed the workers are around their central density at a given point in time. The standard distance d is calculated using Equation (28).

$$d = \sqrt{\frac{\sum_{i=1}^n (j_i - \mu_j)^2 + \sum_{i=1}^n (k_i - \mu_k)^2}{n}} \quad (28)$$

d. Standard Deviational Ellipse:

Shows the directional effect of dispersion. The standard ellipse in this case indicates the directional dispersion of workers by calculating the standard distances for the Easting and Northing axes. The center of the ellipse is the mean center, and the major elliptical axis follows the direction of major workers' dispersion. This is highly useful as it indicates the directional dispersion of workers as they move throughout the site. Thus, indicating the locations of the areas on site where the least workers' density occurs and could indicate the regular workflow of the workers on site.

The standard deviation ellipse can be calculated by assigning weights to different points based on their proximity to one another. The rotated semi-major (σ_j) and semi-minor (σ_k) axes of a weighted directional distribution can be calculated as follows:

$$\begin{pmatrix} \tilde{j} \\ \tilde{k} \end{pmatrix} = w \cdot \begin{pmatrix} j_i \\ k_i \end{pmatrix} - \begin{pmatrix} \mu_j \\ \mu_k \end{pmatrix} \quad (29)$$

$$\sigma_j = \sqrt{\frac{1}{n} \sum_1^n (\tilde{j}_i \sin \theta + \tilde{k}_i \cos \theta)^2} \quad (30)$$

$$\sigma_k = \sqrt{\frac{1}{n} \sum_1^n (\tilde{k}_i \cos \theta + \tilde{j}_i \sin \theta)^2} \quad (31)$$

Where w is the weight matrix and the rotation angle θ is calculated using Equation (32).

$$\tan \theta = \frac{(\sum_{i=1}^n \tilde{j}_i^2 - \sum_{i=1}^n \tilde{k}_i^2) + \sqrt{(\sum_{i=1}^n \tilde{j}_i^2 - \sum_{i=1}^n \tilde{k}_i^2)^2 + 4 \sum_{i=1}^n \tilde{j}_i \tilde{k}_i}}{2 \sum_{i=1}^n \tilde{j}_i \tilde{k}_i} \quad (32)$$

3.3.2.3.2 Spatial Randomness and Clustering of Workers on Site

One way of defining the spatial randomness or clustering of Workers on site is to use the nearest neighbor distance analysis. The analysis indicates whether the point pattern of workers is completely random or if there are clusters of workers centered around different mean centers. In order to implement the nearest neighbor analysis, firstly, the mean of NND, (μ_{NND}), of the workers' is calculated from Equation (33).

$$\mu_{NND} = \frac{\sum_{i=1}^n d_i}{n} \quad (33)$$

Where d_i is the NND of a worker with their nearest neighboring workers. Then the expected mean NND of workers in complete spatial randomness is calculated as:

$$\mu_{CSR} = \frac{0.5}{\sqrt{\frac{n}{A}}} \quad (34)$$

Where A is the area of the minimum bounding box of the workers' coordinates. The ratio between μ_{NND} and μ_{CSR} indicates the distribution of the workers, if the ratio is < 1 then the workers are dispersed, if the ratio is > 1 then the workers are clustered. Finally, the z-score of the mean NND is calculated as:

$$z = \frac{\mu_{NND} - \mu_{CSR}}{DE} \quad (35)$$

Where DE is obtained using Equation (36)

$$DE = \frac{0.261356}{\sqrt{\frac{n^2}{A}}} \quad (36)$$

The z-score then indicates the level of confidence of the workers distribution on a construction site. The higher the z-score the more significant the pattern is.

Also, the G-distance function of the workers' coordinates is used to indicate the clustering or dispersion pattern of the workers at different locations on site. G for workers is defined as:

$$G(d) = \frac{\sum(D_{jk} < d)}{n} \quad (37)$$

Where D_{jk} is the minimum distance between neighboring workers that is less than distance d used to develop the G function for the workers on site. G is calculated for a cumulative d from zero to the maximum distance possible between workers, which reflects the bounding box of the workers' coordinates. The G function of workers following a complete spatial random point pattern, $Gg(d)$ is plotted and the results from $G(d)$ are compared against it to determine the clustering or dispersion of the workers on site.

Another way to test the workers' point pattern is to attempt to define workers' clusters found within a construction site using BIRCH clustering algorithms. It is important to note that, if the workers' density follow complete spatial randomness, detected clustered will not match with any clustering criteria and the results will be misleading.

Using the n number of workers coordinates, and the number of clusters K , clusters are distinguished using Equations (17), (18), (19), and (20).

Knowing the dispersion or clustering of workers on a construction site has potential benefits. If workers are clustered, this means areas where clustering is identified are areas of high activity and are expected to have high productivity. If workers are dispersed, this could potentially mean that most of the workers are idle, and this could indicate low progress on site. The different clusters could also indicate workers working within similar trades, thus indicating on-going activities on site. The result could then be compared against the projects' baseline schedule to identify if the project is on schedule or is experiencing unrecognized delays. Also, by identifying the clusters on site, these clusters can be matched against pre-defined data to ensure that the site is behaving in a traditional manner according to the schedule and on-going activities. This could mean that if clusters are defined by nearest distances, workers from the same crews are assumed to be identified within a cluster. Irregularities in site behavior could be detected if the assumption is false.

3.3.2.3.3 *Workers' Density on Site*

Workers' Density signifies the locations on site at any given point in time, where the highest concentration of workers takes place. The areas of higher densities could be translated into areas of high activity. Such densities could be compared against other site performance parameters to indicate whether there exists a correlation between workers' spatial concentration and how the site is operating. There are three techniques that could calculate and visualize workers' density:

- a. Quadrat Analysis
- b. Voronoi-based analysis
- c. Kernel Density Estimation – KDE as explained earlier in this chapter

a. Quadrat Density:

Using quadrat density analysis, the spatial randomness and clustering of the workers on a construction site can be examined. The site area is divided into sub-regions and the

point density of workers is calculated within each sub-region. The point density is obtained by counting the number of workers' (j, k) coordinates, within each quadrat.

The null hypothesis is that the workers' follow a complete spatial random distribution. Then, a p -value less than 0.05 means that the densities of the workers follow a more clustered or uniformly dispersed behavior rather than a complete spatial random one, and vice versa. t value for workers' density is calculated using Equation (38).

$$t = \frac{\frac{\sum d}{n}}{\sqrt{\frac{\sum d^2 - (\frac{\sum d}{n})^2}{(n-1)(n)}}} \quad (38)$$

where d is calculated as follows,

$$d = \sqrt{(j_i - j_{i+n})^2 + (k_i - k_{i+n})^2} \quad (39)$$

The p -value for the workers' quadrat density is acquired from the normal distribution graph in Figure 3-17. Results from this technique are not very reliable when looking at spatial randomness of workers, however, is an acceptable indicator of workers' densities within the desired site quadrats.

b. Voronoi-based Analysis:

Voronoi-based analysis is another measure of workers' densities on site. The Voronoi generators represent the workers' coordinates. R_k is defined for each worker using Equation (40).

$$R_k = \{g \in \mathbb{R}^2 \mid |g - c_i| \leq |g - c_l| \text{ for all } c \in C\} \quad (40)$$

Where P is the set of workers' coordinates $\{p_1 = (j_1, k_1), \dots, p_n = (j_n, k_n)\}$ in \mathbb{R}^2 , and q is the generator.

Given that the region or cell of a generator shows the space of least distances to the boundaries of the closest generators, this means that the small the region, the more people surrounding the worker. Hence, an area of high workers' density will have smaller polygon regions around each worker, then an area of low workers' density. Also, the Voronoi region of each worker could also signify the working space that the

worker has at a given point in time. The working space would signify the degree of freedom a worker has to execute the required task, and the degree of freedom could be used to determine the working element. An example of that would be, if a crew of steel fixers have small neighboring Voronoi cells this could mean that the workers are fixing the steel reinforcement of a vertical element since it is expected that all workers would be crowded together around the column. Nevertheless, if the workers have larger Voronoi cells this could indicate that a horizontal element is being executed since there is a larger area of work that may be occupied.

c. Kernel Density Estimation – KDE:

KDE can be used to estimate the workers' spatial density over a period of time. By determining the frequency of occurrence of a workers' (j, k) coordinate within a certain location on site, the density function of workers' can be plotted as shown in Figure 3-20.

By interpreting a KDE contour plot, i.e., heatmap, workers' density functions can be depicted on site. The more workers located in a certain area, the higher the probability density will be in that area, the closer the contour lines will be in a KDE plot, and the higher the intensity of the color is in that area. According to the plot and the heatmap, the level of activity in the different zones can be depicted. Colors representing the location of high activity have a higher intensity than the rest of the areas. The zones on site can be categorized into three types based on the level of activity, these are:

- a. High-Activity Zones – where the high-activity zone density (*HAZD*) is defined by densities falling within the range of $[0.75, 1]$.
- b. Medium-Activity Zones – where the medium-activity zone density (*MAZD*) is defined by densities falling within the range of $[0.5, 0.75]$.
- c. Low-Activity Zones – where the low-activity zone density (*LAZD*) is defined by densities falling within the range of $[0, 0.5]$.

3.3.2.3.4 *Workers' Time Distribution*

Lastly, after knowing the areas of workers' concentration, the distribution of the workers' time on site can be determined. By analyzing the location of workers in an area on site, it can be implied whether the worker is actually working, traveling, or idle. This is done by identifying the bounding zone in which the worker's coordinate is located. In other words, if the coordinate of the worker is located in working area, then the worker is assumed to be working. If the worker is located in a resting area, then the worker is assumed to be idle, and if the is located on site but does not exist in any of the defined areas, then the worker is said to be traveling. Finally, the percentage of working time, idle time, and traveling time for each worker or all workers on site can be identified. The percentages can indicate the productivity of the workers on site, higher percentages of working time, usually translate to higher productivity, and vice versa. If workers spend most of their time traveling on site, this could indicate a problem with the sites' layout, and could mean that there is time wasted which could be utilized had the site layout been enhanced. The percentages could be calculated using the logic flowchart in Figure 3-26.

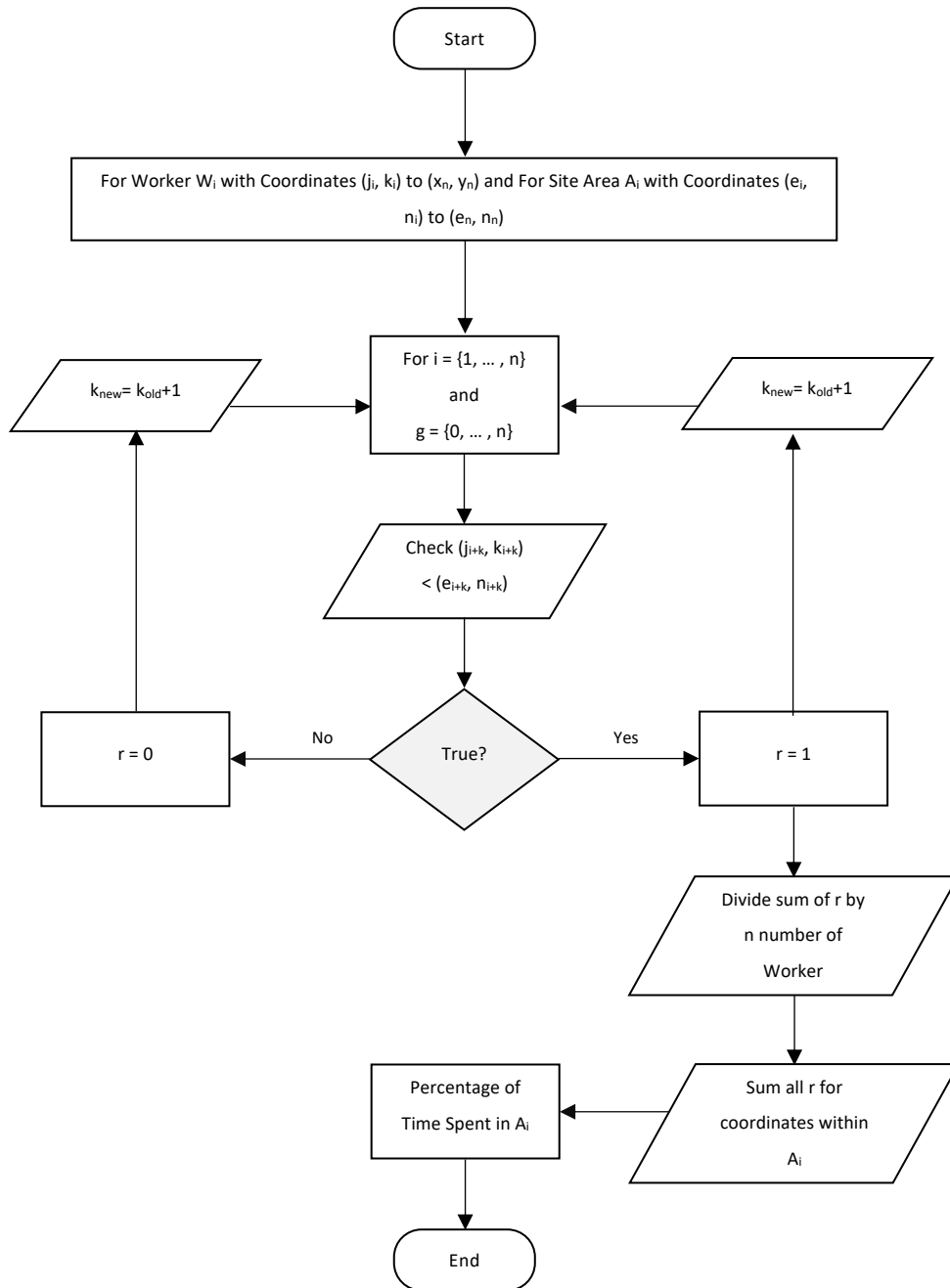


Figure 3-26: Flowchart for Workers' Time Distribution Calculation

After the above logic is iterative on all workers' coordinates, and the final output is shown as in Table 3-10, where the summation of the percentages must equal to 1.

Table 3-10: Output of Workers' Time Distribution

Time Category	%
Working	X
Traveling	Y
Idle	Z

A major assumption made when calculating the above percentages is that the location of the worker is the only measure upon which the categorization has been made. This means that an idle worker within a construction area cannot be detected. However, the percentages are still a satisfactory mean of identifying how a worker's time is spent on site.

To provide a recap of all the stages, methods, and outputs discussed under this section, a summary of the entire first order analysis framework is shown in Figure 3-27.

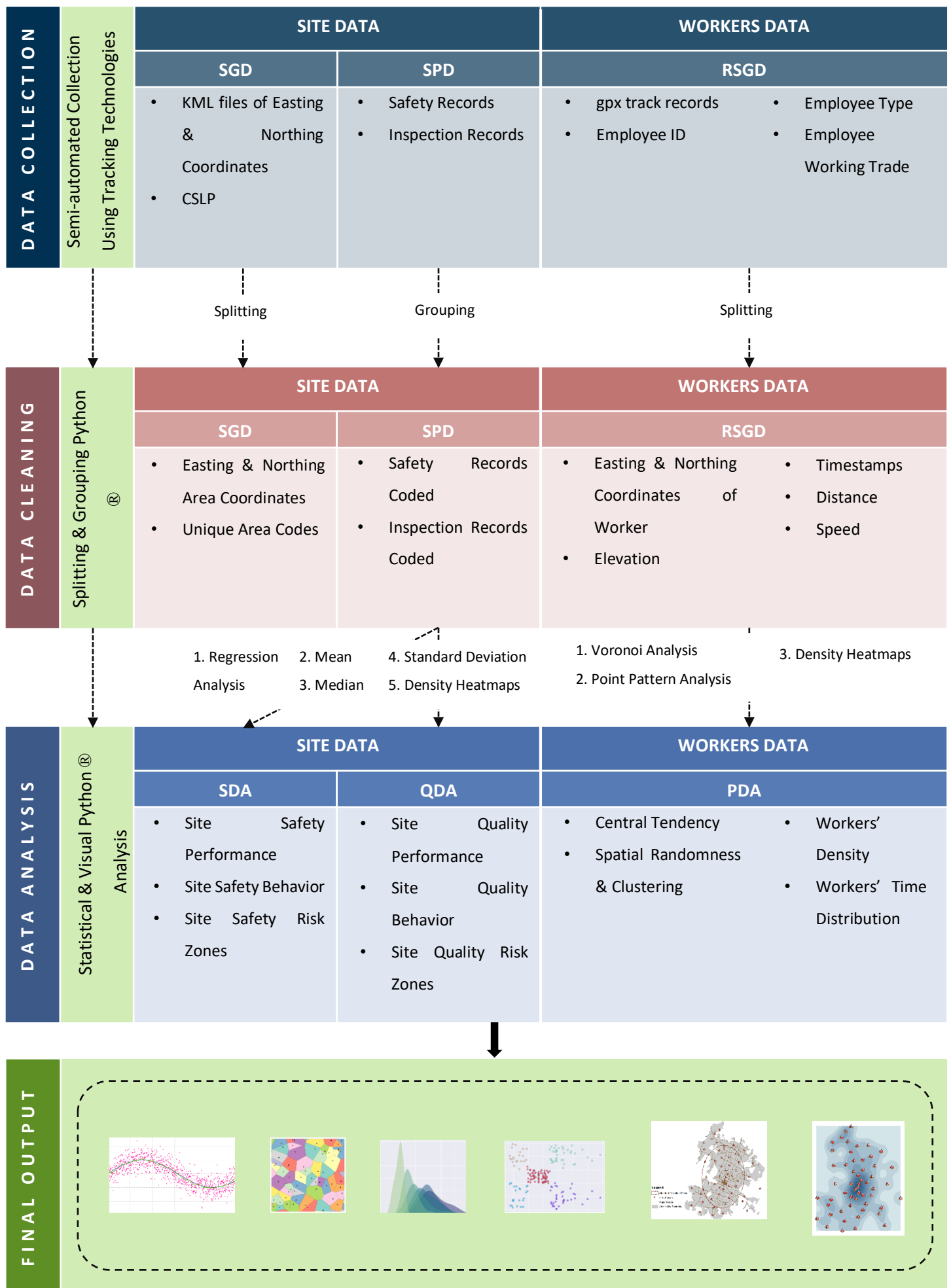


Figure 3-27: Detailed First Order Analysis Framework Composition

CHAPTER 4 – FRAMEWORK IMPLEMENTATION FOR SECOND ORDER ANALYSIS

Outputs of the first order data analysis (*FODA*) obtained under section 3.3.2 – *Analysis Techniques Implementation and Outputs*, could be used to generate a second order analysis outputs (*SOAO*). Namely, when the results of the first order analysis are compared against one another, second order correlative data analysis (*SOCDA*) could be performed. For the sake of demonstrating how the correlative analysis could be done, the *FODA* is implemented on a set of randomly generated site data. Results from the *FODA* are then integrated to reach the *SOAO*. Accordingly, this chapter covers the following:

1. First Order Data Analysis (*FODA*) – Analysis done separately as explained under section 3.3.2 – *Analysis Techniques Implementation and Output*:
 - a. *SDA*
 - b. *QDA*
 - c. *PDA*
2. Second Order Correlative Data Analysis (*SOCDA*) – Semi-visual and statistical correlative analysis applied on outputs obtained from the *FODA*. The purpose of the *SOCDA* is to attempt to depict the existence of relationships between the different site performance parameters. The outputs from the *FODA* are used for *SOCDA* as per the matrix shown in Table 4-1.

Table 4-1: Second Order Analysis Matrix

		P R O D U C T I V I T Y				
		Spatial Temporal Location	Central Tendency	Density	Clustering	Distribution
S A F E T Y	Mean & Median				Correlation	Correlation
	Moving Average	Correlation				
	Density	Correlation	Correlation	Correlation		
Q U A L I T Y	Mean & Median				Correlation	Correlation
	Moving Average	Correlation				
	Density	Correlation	Correlation	Correlation		
O T H E R S	Site Layout		Efficiency	Efficiency		Efficiency
	Schedule	Progress Variance		Progress Variance		
	Cost	Value for Money		Value for Money		Value for Money

The demonstration process is broken down into 3 stages to reach the final *SOCDA*, these stages are; (1) data generation and visualization, where random site data for *SGD*, *SPD*, and *RGSD* is generated for the purpose of implementation and not for verification. The data is then plotted using scatterplots for the aim of spatial visualization and description, (2) first order data analysis (*FODA*), where the framework for first order data analysis is applied on the random data, and the outputs are used for the purpose of achieving the, (3) second order correlative analysis (*SOCDA*), from which second order analysis outputs (*SOAO*) are obtained.

4.1 Stage 1 - Data Generation and Visualization

Firstly, using Microsoft 365 - Microsoft ® Excel ® 2013, random data was generated using the *RANDBETWEEN* function to produce random numbers for the site

geographic data, for the site periodical data, and the real-time geospatial data of site workers. The visualization was then done using Python ®.

4.1.1 Site Geographic Data – (SGD)

The chosen site has a rectangular boundary of minimum and maximum coordinates (0,0) and (20,30), respectively. Then the site is assumed to have a layout comprised of three construction zones, one main caravan, one resting area, one workshop, and one storage area. The locations and areas of the zones are obtained randomly, and the SGD is then constructed shown in Table 4-2.

Table 4-2: Randomly Generated SGD

Area_Code	Ref_No.	Easting	Northing	Elevation	Category
C01	1	10.00	11.00	412.00	WA
C01	2	7.00	11.00	412.00	WA
C01	3	7.00	14.00	412.00	WA
C01	4	10.00	14.00	412.00	WA
C02	5	2.00	2.00	413.00	WA
C02	6	2.00	7.00	413.00	WA
C02	7	8.00	2.00	413.00	WA
C02	8	8.00	7.00	413.00	WA
C03	9	3.00	11.00	415.00	WA
C03	10	1.00	11.00	415.00	WA
C03	11	1.00	15.00	415.00	WA
C03	12	3.00	15.00	415.00	WA
MA01	13	16.00	16.00	412.00	WA
MA01	14	16.00	19.00	412.00	WA
MA01	15	19.00	19.00	412.00	WA
MA01	16	19.00	16.00	412.00	WA
R01	17	16.00	2.00	413.00	RA
R01	18	16.00	4.00	413.00	RA
R01	19	18.00	4.00	413.00	RA
R01	20	18.00	2.00	413.00	RA
W01	21	10.00	1.00	414.00	WA
W01	22	10.00	5.00	414.00	WA
W01	23	14.00	5.00	414.00	WA
W01	24	14.00	1.00	414.00	WA
S01	25	1.00	20.00	415.00	WA
S01	26	1.00	24.00	415.00	WA
S01	27	12.00	24.00	415.00	WA

S01	28	12.00	20.00	415.00	WA
B	29	0.00	0.00	412.00	TA
B	30	0.00	30.00	412.00	TA
B	31	20.00	30.00	412.00	TA
B	32	20.00	0.00	412.00	TA

The data is generated through choosing a random point (10,11) for C01. Then the list of lengths and widths shown in Table 4-3 is used to generate the site layout. Figure 4-1 shows the final plot of the predetermined site areas, where all site areas were assumed to be rectangular in shape.

Table 4-3: Dimensions of Site Areas in meters

Area_Code	Length (m)	Width (m)
C01	3	3
C02	6	5
C03	2	4
MC01	3	3
R01	2	2
W01	4	4
S01	11	4
B	20	30

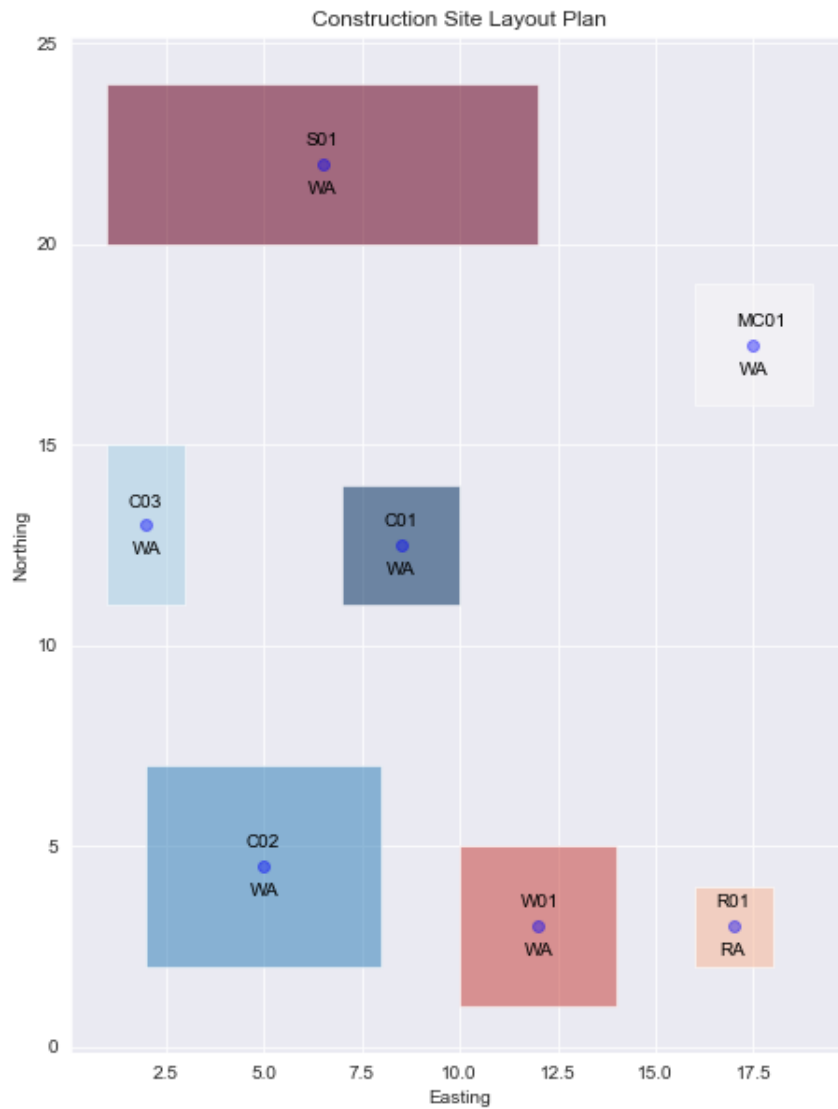


Figure 4-1: Construction Site Layout for Random SGD

The site layout plot shows the site zones as filled polygons. Each zone is represented by a different color, the centroids of the zones are shown on the plot as well, alongside the labels for the Area Code and the Category of the area.

4.1.2 Site Periodic Data – (SPD)

Using the same RANDBETWEEN function, data for the safety and inspection records on site are randomly obtained. The time period over which the data is assumed to be collected and analyzed is 15 days, hence t for analysis is equal to 15, from 15th of December 2020 till 29th of December 2020. The working hours on site are assumed to be normal working hours from 8:00:00 till 17:00:00.

4.1.2.1 Safety Records

The number, the location, date, and time of safety incidents on the chosen site are produced as shown in Table 4-4. The maximum and minimum number of incidents are 0, and 10, respectively.

Table 4-4: Randomly Generated SPD - Safety Records

Area_Code	No_Incidents	Date
C01	5	12/15/20
C02	6	12/16/20
C03	3	12/17/20
C01	8	12/18/20
C02	7	12/19/20
C03	5	12/20/20
C01	3	12/21/20
C02	8	12/22/20
C03	4	12/23/20
C01	5	12/24/20
C02	3	12/25/20
C03	2	12/26/20
C01	5	12/27/20
C02	4	12/28/20
C01	7	12/29/20

4.1.2.2 Inspection Records

The number, the location, date, and time of accepted and rejected inspection records on the chosen site are produced as shown in Table 4-5. The maximum and minimum number of accepted inspection records are 0, and 10, respectively. Also, the maximum and minimum number of rejected inspection records are 0, and 15, respectively.

Table 4-5: Randomly Generated SPD - Inspection Requests

Area_Code	Rejected	Accepted	Date
C01	10	2	12/15/20
C02	9	0	12/16/20
C03	2	3	12/17/20

C01	1	4	12/18/20
C02	3	3	12/19/20
C03	9	5	12/20/20
C01	10	3	12/21/20
C02	12	8	12/22/20
C03	4	7	12/23/20
C01	4	5	12/24/20
C02	4	3	12/25/20
C03	5	6	12/26/20
C01	7	0	12/27/20
C02	8	4	12/28/20
C01	10	1	12/29/20

4.1.3 Real-time Geospatial Data – (RGSD)

The construction site is assumed to have three workers, two of them are skilled laborers and one is a junior site engineer. The skilled labors are presumed to be a concrete pouring labor, a formwork and a scaffolding labor. Workers geospatial and temporal data are randomly generated between the site boundary coordinates and time period t , respectively. The data shall follow complete spatial randomness for the purpose of implementation verification. The produced data is shown in Table 4-6.

Table 4-6: Sample of Randomly Generated RGSD

Employee_ID	Employee_Type	Working_Activity	Easting	Northing	Elevation	Distance	Speed	Time	Date
E01	Skilled Labor	Concrete	6.26	29.32	416	518	1.82	8:00 AM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	4.50	23.81	414	794	1.82	8:00 AM	12/15/20
E03	Junior Engineer	Supervision	18.38	26.94	414	55	0.91	8:00 AM	12/15/20
E01	Skilled Labor	Concrete	6.52	13.48	416	650	0.91	9:00 AM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	15.37	9.27	415	556	1.82	9:00 AM	12/15/20
E03	Junior Engineer	Supervision	2.77	23.98	412	275	0.00	9:00 AM	12/15/20
E01	Skilled Labor	Concrete	8.04	9.27	415	730	0.91	10:00 AM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	15.59	18.89	416	443	2.73	10:00 AM	12/15/20
E03	Junior Engineer	Supervision	12.49	15.90	415	328	0.91	10:00 AM	12/15/20
E01	Skilled Labor	Concrete	2.03	24.05	415	348	1.82	11:00 AM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	7.84	24.49	414	472	0.00	11:00 AM	12/15/20
E03	Junior Engineer	Supervision	5.45	16.51	415	0	2.73	11:00 AM	12/15/20
E01	Skilled Labor	Concrete	18.16	27.02	416	443	1.82	12:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	12.68	23.50	413	96	2.73	12:00 PM	12/15/20
E03	Junior Engineer	Supervision	4.67	27.79	416	178	0.91	12:00 PM	12/15/20

E01	Skilled Labor	Concrete	6.23	13.09	416	844	1.82	1:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	13.83	24.13	413	742	1.82	1:00 PM	12/15/20
E03	Junior Engineer	Supervision	10.32	3.99	412	806	0.00	1:00 PM	12/15/20
E01	Skilled Labor	Concrete	13.90	25.56	413	517	0.00	2:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	19.77	11.11	413	478	0.91	2:00 PM	12/15/20
E03	Junior Engineer	Supervision	14.37	29.14	412	508	0.91	2:00 PM	12/15/20
E01	Skilled Labor	Concrete	14.78	26.12	415	965	0.00	3:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	18.75	2.47	412	444	1.82	3:00 PM	12/15/20
E03	Junior Engineer	Supervision	7.70	27.83	416	905	0.91	3:00 PM	12/15/20
E01	Skilled Labor	Concrete	13.04	13.97	416	527	0.91	4:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	2.67	20.94	416	934	2.73	4:00 PM	12/15/20
E03	Junior Engineer	Supervision	11.24	16.59	416	882	0.91	4:00 PM	12/15/20
E01	Skilled Labor	Concrete	6.23	24.44	415	342	0.00	5:00 PM	12/15/20
E02	Skilled Labor	Formwork & Scaffolding	16.57	7.37	412	798	0.91	5:00 PM	12/15/20
E03	Junior Engineer	Supervision	17.18	16.71	413	959	1.82	5:00 PM	12/15/20
E01	Skilled Labor	Concrete	14.72	19.55	413	332	0.00	8:00 AM	12/16/20
E02	Skilled Labor	Formwork & Scaffolding	9.49	7.07	415	843	0.91	8:00 AM	12/16/20
E03	Junior Engineer	Supervision	3.00	2.78	415	88	0.91	8:00 AM	12/16/20
E01	Skilled Labor	Concrete	11.88	7.44	414	825	0.00	9:00 AM	12/16/20

Coordinates from the dataset is shown on a scatterplot for visualization of the workers' location. Figure 4-2 shows the locations of the workers during the entire day on the 23rd of December 2020. The workers are grouped by the type of activity they are working on for enhanced monitoring and analysis. A timeseries bar is also added to allow for continuous visualization of the workers' location on different days during the project's duration.

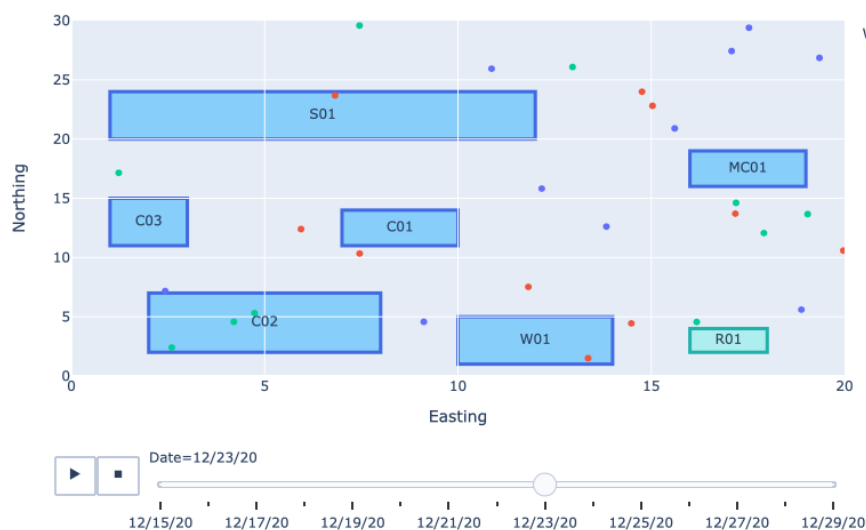


Figure 4-2: Scatterplot of Workers' Coordinates

The scatterplot shown in Figure 4-3 displays the same data in Figure 4-2, on the 15th of December 2020, however, the visualization windows are split for each working activity, as this allows for improved real-time monitoring of each working crew on site.



Figure 4-3: Scatterplot of Aggregated Workers' Coordinates

Finally, the coordinates of the workers are connected in the spatial plane to show the daily tracks of the workers' as they travel around on site. A sample of the daily tracks are shown in Figure 4-4.



Figure 4-4: Line plot of Aggregated Workers' Daily Tracks

4.2 Stage 2 – First Order Data Analysis (*FODA*)

After the data is generated, is it analyzed using the first order analysis techniques in the same manner discussed under section 3.3.2 – *Analysis Techniques Implementation and Outputs*. The outputs from implementing the techniques on the generated data are explained and interpreted under this section.

4.2.1.1 Safety Data Analysis (SDA)

For safety data analysis the techniques are used to assess the random data for safety incidents on site accordingly:

1. Site Safety Performance (SSP) which is provided by:
 - a. *Temporal Mean.*
 - b. *Temporal Median.*
 - c. *Temporal Standard Deviation.*
2. Site Safety Behavior (SSB) which is provided by:
 - a. *Moving Average Linear Regression.*
3. Site Safety Risk Zones (SSRZ) using heatmaps that show:
 - a. *Safety High-risk Zones*
 - b. *Safety Medium-risk Zones*
 - c. *Safety Low-Risk Zones*

4.2.1.1.1 Site Safety Performance (SSP)

Firstly, the mean, median and standard deviation of the number safety incidents, over a time period $t = 15$ days, on site are calculated. The results used to assess the SSP are shown in Figure 4-5.

No_Incidents	
max	8.000000
min	2.000000
mean	5.000000
median	5.000000
std	1.889822

Figure 4-5: Mean, Median, and Standard Deviation of Safety Incidents on Site

a. Temporal Mean:

The average number of safety incidents on the site is **5 incidents per day**. When compared with the construction industry norm, if this mean is higher than the norm, then the site has poor safety standards. If the mean is less than the industry norm, then the site has satisfactory to high safety standards.

b. Temporal Median:

The median of the number of safety incidents is also equal to **5 incidents per day**, which is similar to the mean. This means that the site does not encounter an outlying number of safety incidents over the 15 working days. In the case of outliers, the median will be more accurate than the mean, and in this case, the median is the measure compared against the construction industry norm instead of the mean.

c. Standard Deviation:

The standard deviation of **1.889 incidents**, rounded up to **2 incidents**, means distribution of the incidents on site cover a range of **5 ± 2 incidents**.

The lower the standard deviation is, the more consistent the safety conditions are on site, and vice versa.

4.2.1.1.2 Site Safety Behavior (SSB)

a. Moving Average Linear Regression:

The moving average of the safety incidents was calculated using a window of **3 days** since $t = 15$ days. The plot of the 3-day moving average and the daily number of safety incidents on site is shown in Figure 4-6.

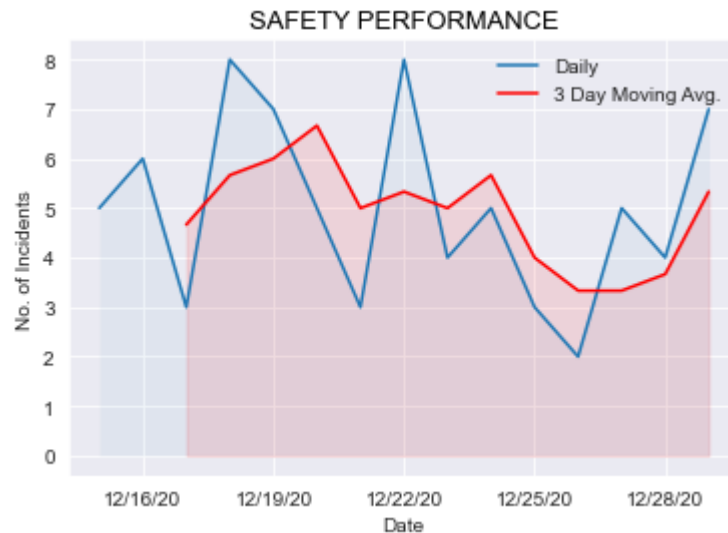


Figure 4-6: Moving Average Regression of Safety Incidents on Site

The x-axis represents the days, and the y-axis represents the number of safety incidents. The moving average plot shown above does not depict a constant or specific behavior for the incidents over the 15-day period, since the average increases and decreases without displaying an identifiable pattern. However, if an identifiable relationship is depicted, this can be used to predict future behavior of the site in terms of safety performance.

4.2.1.1.3 Site Safety Risk Zones (SSRZ)

a. Kernel Density Estimation – KDE:

Using the KDE analysis, the probability density of the number of safety accidents can be obtained and presented as shown in Figure 4-7.

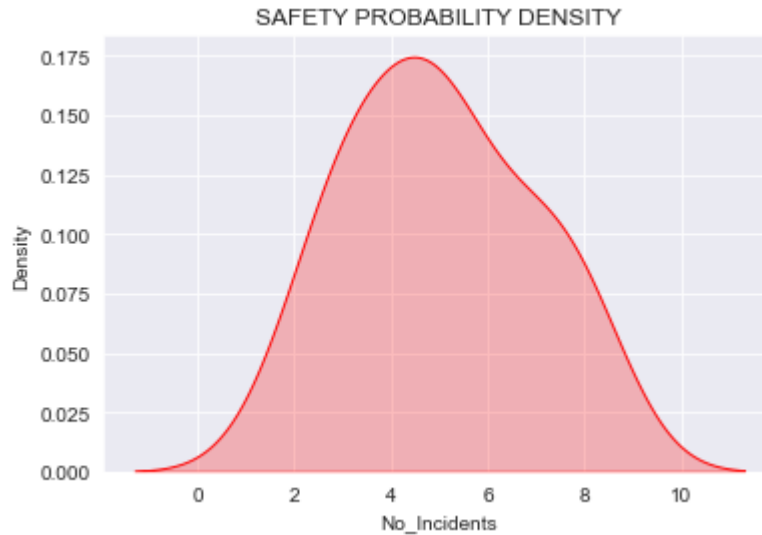


Figure 4-7: Probability Density Plot of Safety Incidents on Site

The x-axis of the plot represents the number of safety incidents, and the y-axis represents the probability density of the incidents. The plot indicates that the highest frequency of safety incidents occurring over the 15-day time period on site is between **3 and 5 incidents**.

Using the KDE plot and knowing the location of each of the incidents, the incidents are assigned to the centroids of the areas shown previously in Figure 4-1, and the heatmap of the safety incidents on site is obtained in the Easting-Northing Spatial Plane.

The heatmap shows the locations of safety risk zones on site as seen in Figure 4-8.

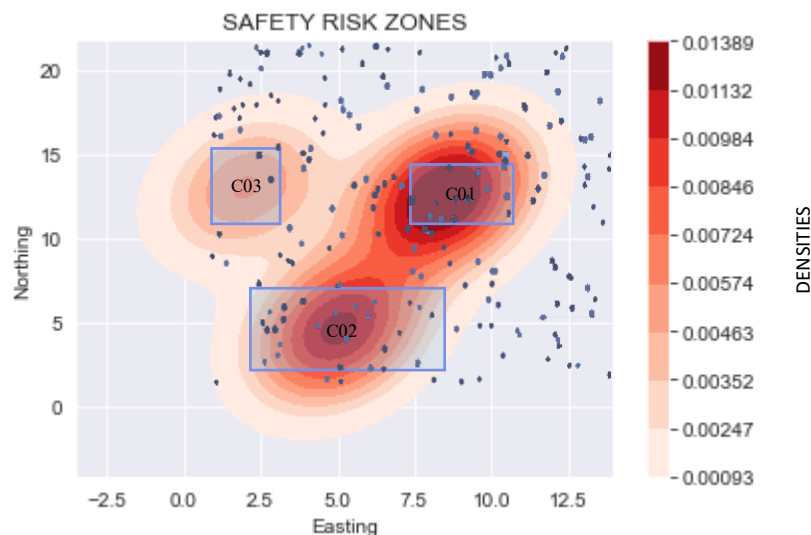


Figure 4-8: Safety Risk Zones - Heatmap of Safety Incidents on Site

The areas of high color intensity indicate the areas of safety high-risk zones, these are the areas of the highest probability density of safety incidents on site. As the intensity becomes less, the areas experience lower densities of incidents thus are classified as either medium-risk or low-risk zones depending on the gradient of the color bar. The gradient of the color bar is determined according to the probability distribution function as shown in the heatmap plot.

4.2.1.2 Quality Data Analysis (QDA)

4.2.1.2.1 Site Quality Performance (SQP)

Firstly, the mean, median and standard deviation of the number both accepted and rejected inspection requests, over a time period $t = 15$ days, on site are calculated. Also, the maximum and minimum number of requests alongside other information about the location, time, and date on which these incidents occurred are extracted. The final results used to assess the SQP are shown in Figure 4-9.

	Rejected	Accepted
max	12.000000	8.000000
min	1.000000	0.000000
mean	6.533333	3.600000
median	7.000000	3.000000
std	3.440653	2.354327

Figure 4-9: Mean, Median, and Standard Deviation of Inspection Requests on Site

d. Temporal Mean:

The average number of accepted and rejected requests on the site are **4 and 7 requests per day**, respectively. Both means should be compared against each other to assess the quality performance on site. Ideally, the average of the rejected inspection requests should be zero. However, given the nature of the construction field and the existing probability of faulty execution, the average of the rejected inspection requests should be 60% less than the average of the accepted inspection requests, for a site that has high quality of execution.

e. Temporal Median:

The median of the number of accepted and rejected inspection requests is equal to **3 and 7 incidents per day**, respectively, which are similar to the means. This signifies that the site does not encounter an outlying number of requests over the 15 working days. In the case of outliers, the median will be more accurate than the mean, and in this case, the median is the measure used to indicate the quality performance on site rather than the mean.

f. Standard Deviation:

The standard deviation of **3 rejected requests and 2 accepted requests**, means distribution of the requests on site cover a range of **7 ± 3 requests and 3 ± 2 requests** for rejected and accepted requests respectively.

The lower the standard deviation is, the more consistent the quality of execution is on site, and vice versa.

4.2.1.2.2 Site Quality Behavior (SQB)

b. Moving Average Linear Regression:

The moving average of the requests was calculated using a window of **3 days** since $t = 15$ days. The plot of the 3-day moving average and the daily number of accepted and rejected inspection requests on site are shown in Figure 4-10.

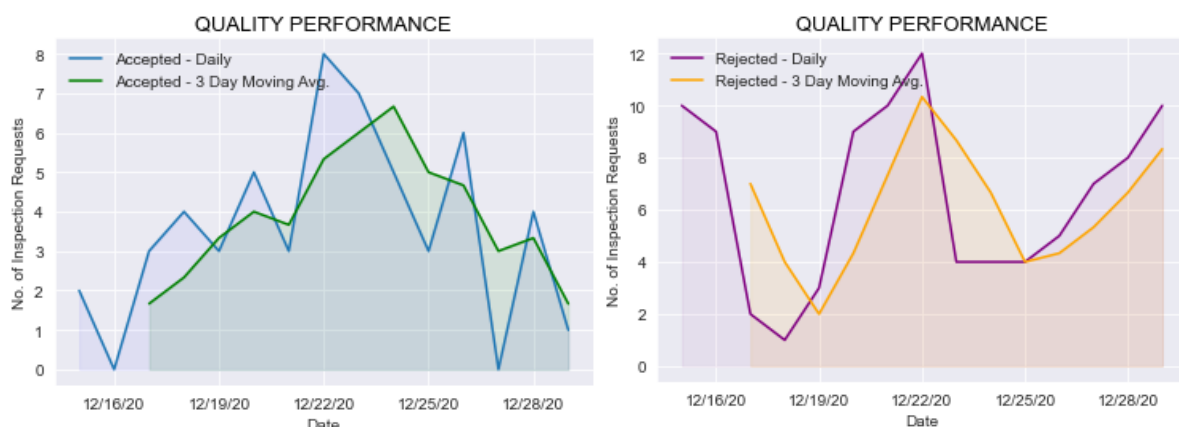


Figure 4-10: Moving Average Linear Regression of Inspection Requests on Site

The x-axis represents the days, and the y-axis represents the number of inspection requests. The moving average plot shown above does not depict a constant or specific

behavior for the requests over the 15-day period, since the average increases and decreases without displaying an identifiable pattern. However, if an identifiable relationship is depicted, i.e., the regression curve is linearly or exponentially increasing or decreasing over time, this can be used to predict future behavior of the site in terms of safety performance.

4.2.1.2.3 Site Quality Zones (SQZ)

b. Kernel Density Estimation – KDE:

Using the KDE analysis, the probability density of the number of inspection requests can be obtained and presented as shown in Figure 4-11.



Figure 4-11: Probability Density Plot of Inspection Requests on Site

The x-axis of the plot represents the number of inspection requests, and the y-axis represents the probability density of the requests. The plot indicates that the highest frequencies of accepted and rejected requests occurring over the 15-day time period on site are between **2 and 4 accepted requests**, and **4 and 10 incidents rejected requests**.

Using the KDE plot and knowing the location of each of the requests, the requests are assigned to the centroids of the areas shown previously in Figure 4-1, and the heatmap of the inspection requests on site are obtained in the Easting-Northing Spatial Plane.

The heatmap shows the locations of quality zones on site as seen in Figure 4-12. Heatmap with a green contouring is for accepted inspection requests, and the heatmap with the orange contouring is for the rejected inspection requests.

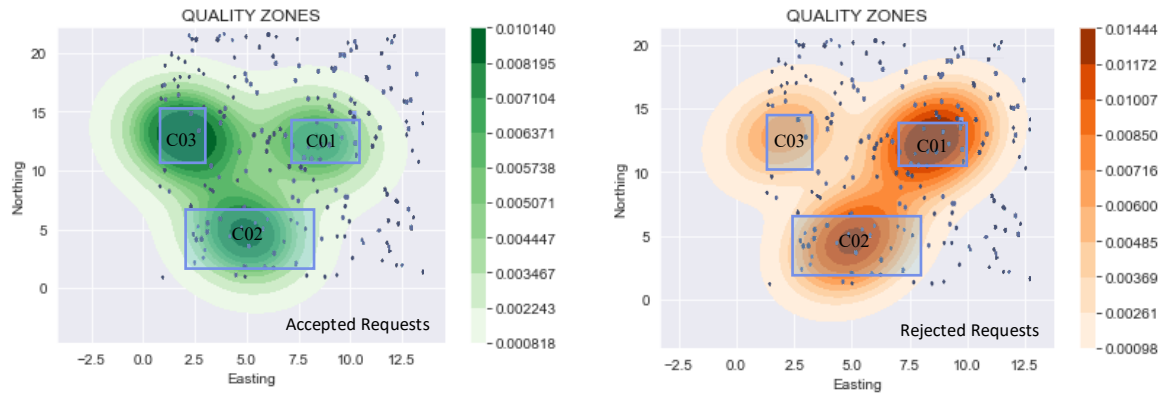


Figure 4-12: Quality Zones - Heatmap of Inspection Requests on Site

The areas of high color intensity, seen on the accepted inspection requests heatmap, indicate the areas of high-quality zones, these are the areas of the highest probability density of accepted inspection requests on site. As the intensity becomes less, the areas experience lower densities of requests thus are classified as either medium-quality or low-quality zones depending on the gradient of the color bar. The gradient of the color bar is determined according to the probability distribution function as shown in the heatmap plot. The opposite is true for rejected inspection requests.

4.2.1.3 Productivity Data Analysis (PDA)

Lastly, the final FODA implemented is used to acquire information regarding the spatial behavior of workers on site using the randomly generated RGSD dataset. Under this section the following will be studied:

1. Central Tendency of Workers on Site
2. Spatial Randomness and Clustering of Workers on Site
3. Workers' Density on Site using Heatmaps that determine:
 - a. High-activity Zones
 - b. Medium-activity Zones
 - c. Low-activity Zones
4. Workers' Time Distribution

4.2.1.3.1 Central Tendency of Workers on Site

Using the PPA methods as explained under section 3.3.3.2.2 - *Productivity Data Analysis (PDA)*, measures could be calculated to determine the central tendencies of workers on the construction site.

a. Mean Center:

Firstly, the mean center of the workers' coordinates is obtained. The mean center is calculated over the working period $t = 15$ days, using the Easting and Northing Coordinates for all workers at all timestamps during the working period. The obtained coordinates of the mean center are shown on a scatter plot which also shows all the obtained workers' coordinates, this is seen in Figure 4-13.

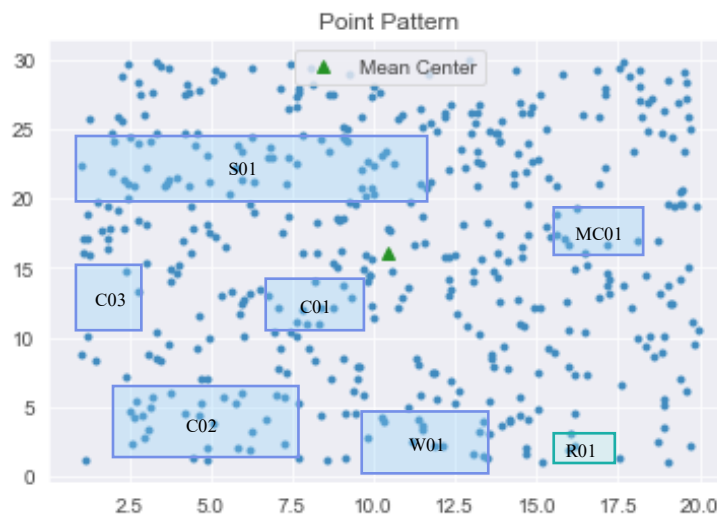


Figure 4-13: Mean Center of Workers on Site

The location of the mean center looks to be at the center of the site boundary, this coincides with the spatial randomness of the data generated.

a. Median Center:

Then, the median center is calculated as it is a more enhanced measure of the central tendencies of workers. Usually, the mean and median center never coincide, as the median center is calculated based on the minimum Euclidean distances between workers whereas the mean is calculated by averaging all the Easting and Northing Coordinates of the Workers. Figure 4-14 shows both the mean and median center of the workers.

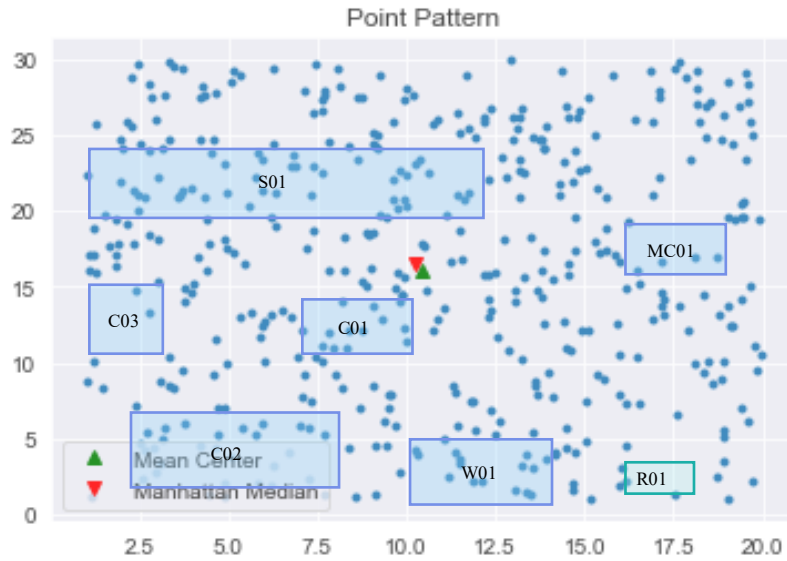


Figure 4-14: Mean Center and Median of Workers on Site

b. Standard Distance:

The standard distance is calculated for the workers on site and is used as the radius for the standard circle encompassing the nearest workers to the mean center. This measure shows the dispersion of workers around the mean center. The higher the standard distance the more dispersed the workers are, the larger the radius of the circle is.

Figure 4-15 shows that the workers are generally dispersed in the North-South axis, this can be explained by the rectangular nature of the site, since the site boundary is longer in the North-South direction.

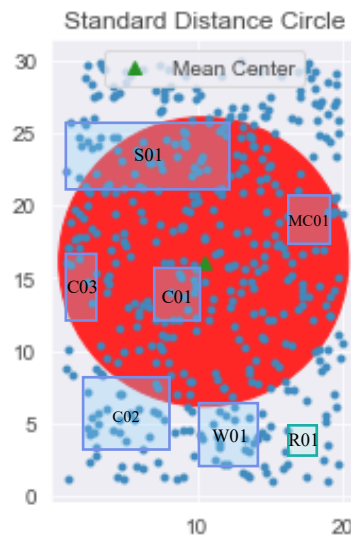


Figure 4-15: Standard Distance Circle of Workers on Site

However, since the radius of the standard circle is almost equal to the East-West length of the site boundary, this means that the workers are dispersed around the mean, thus validating the spatial randomness of the workers on site.

c. Standard Deviational Ellipse:

Moving on, the standard deviational ellipse shows the directional dispersion of workers', giving an indication of the natural flow of workers on site, and highlighting movement routes which the workers usually tend to take. The standard deviational ellipse of the workers is shown in Figure 4-16.

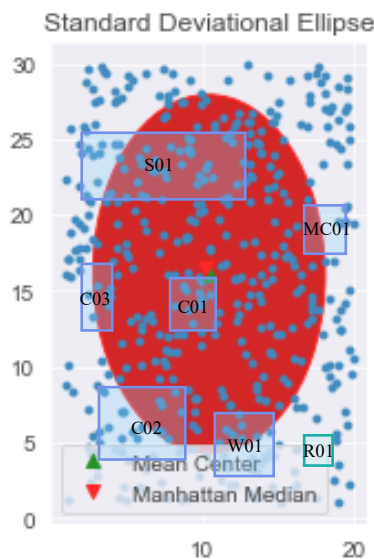


Figure 4-16: Standard Deviational Ellipse of Workers on Site

The ellipse shown again confirms the dispersion of workers as the orientation of the largest axis of the ellipse is parallel to the longest side boundary as explained in the Standard Distance.

4.2.1.3.2 *Spatial Randomness and Clustering of Workers on Site*

The identified measures for workers central tendencies provide a slight indication about the dispersion or clustering of workers on site. However, the NND analysis is used as a more enhanced measure of the dispersion or clustering of workers. The p-value obtained for the NND distance for the workers is **0.547734**, which is greater than 0.05, signifying a random distribution of workers.

A more affirmative indicator of the workers' point pattern is the NND G-distance analysis. The G-distance analysis provides an accurate measure of the randomness or clustering of workers on site. It does so by calculating the densities of the nearest neighboring distances between all workers on site and comparing these densities with the expected densities of a distribution of workers following complete spatial randomness.

Firstly, the G-function is calculated accordingly for a spatially random distribution and plotted against different distances d . Moreover, the G-distances for the workers on site are calculated and shown on the same plot for comparison. If, the G-distance plot, shown in Figure 4-17, for the workers falls below the expected G-distance plot, this means that the densities are observed at larger distance since the curve goes up at a lower rate. In this case, the workers are said to either be uniformly dispersed or randomly distributed. However, If, the G-distance plot for the workers falls above the expected G-distance plot, this means that the densities are observed at smaller distance since the curve goes up at a much higher rate. In this case, the workers are said to either be clustered.

The same assumption can be made using a G-plot envelope, where the G-function is plotted with an envelope of confidence interval 95% as shown in Figure 4-18.

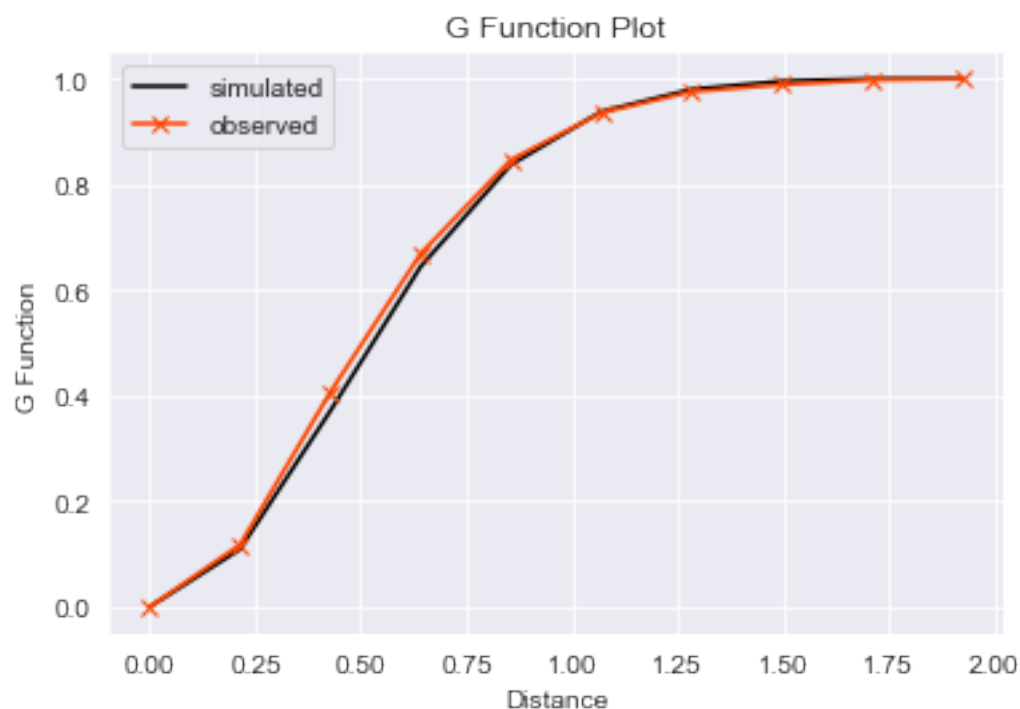


Figure 4-17: G-function Plot for Workers on Site

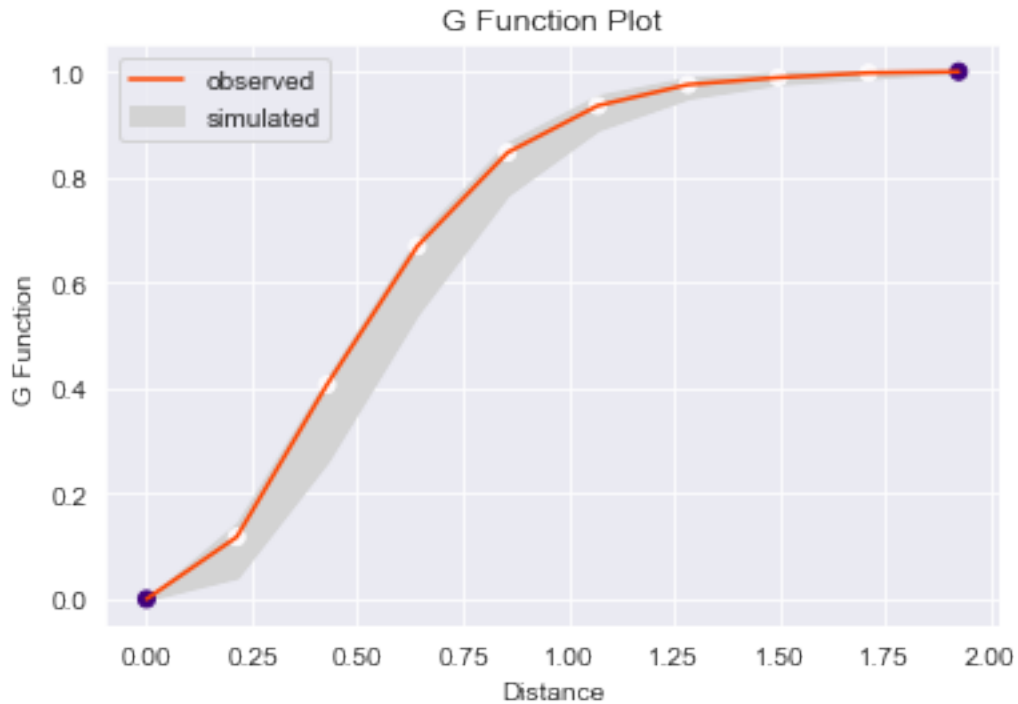


Figure 4-18: G-function Envelope Plot for Workers on Site

In the figures shown, it is observed that the G-distance plot for the observed workers coordinates overlaps exactly with that of a workers' pattern following complete spatial randomness for the given site boundaries. Again, this observation matches with the generated distribution of the workers.

To further analyze the clustering or dispersion of workers, attempts to define workers clusters on site are done using BIRCH Clustering as explained under section 3.3.2.3.2 – *Spatial Randomness and Clustering of Workers on Site*.

The algorithm works iteratively to try and identify clusters based on workers' coordinates on site. Thus, clustering relies heavily on the observed distances between workers, meaning, the closer the workers are in an area, the easier it is to identify these workers as a cluster. It would be expected that on a construction site, workers in the same crew or workers in different crews working on the same activity, would be observed as clusters since they would be assumed to be closer together than other workers on site. Thus, failure to identify clusters on site, indicate spatial randomness of workers.

Given that the data generated for workers coordinates is random, it is probable that no clusters would be identified, or random clusters would be depicted, as shown in Figure 4-19. The 7 detected clusters do not provide any insight on the distances between workers or the working activity of the workers since the workers are randomly distributed on site.

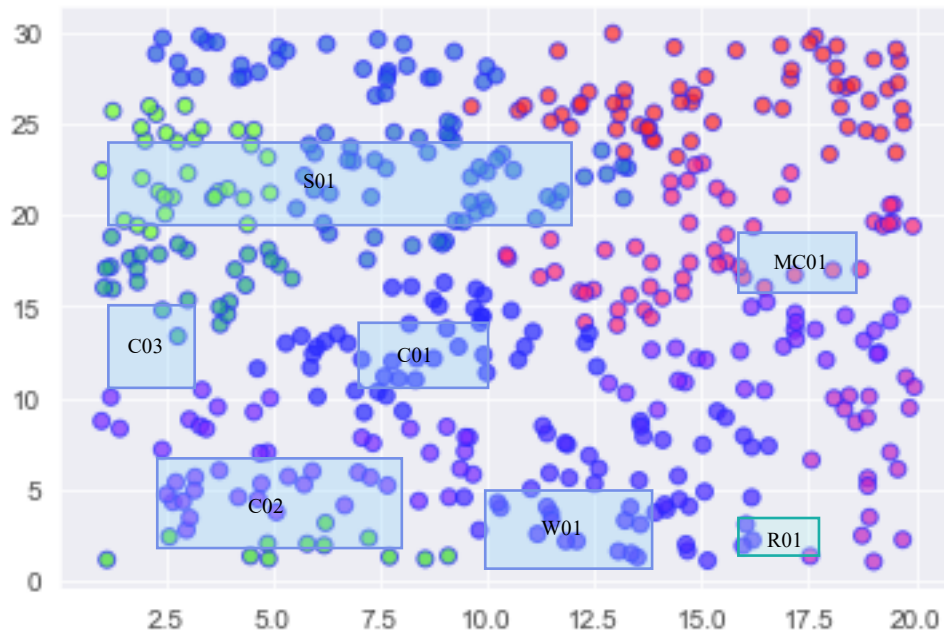


Figure 4-19: BIRCH Clustering of Workers on Site

4.2.1.3.3 Workers' Density on Site

One useful indicator of the workers productivity on site, is the workers' density. Workers' density shows the location of highest and lowest workers concentration. The location of highest workers density on site, would be expected to translate to higher productivity areas, this will be verified later under the Second Order Correlative Analysis. The 3 techniques used to determine the workers density are:

- a. Quadrat Analysis
- b. Voronoi-based analysis
- c. Kernel Density Estimation – KDE

a. Quadrat Density:

Under quadrat density, the site is split into multiple quadrats and the density of workers is calculated in each quadrat. The density is simply the indicator of the count of workers coordinates falling within the quadrat. Results of the quadrat analysis carried out for this site is shown in Figure 4-20.

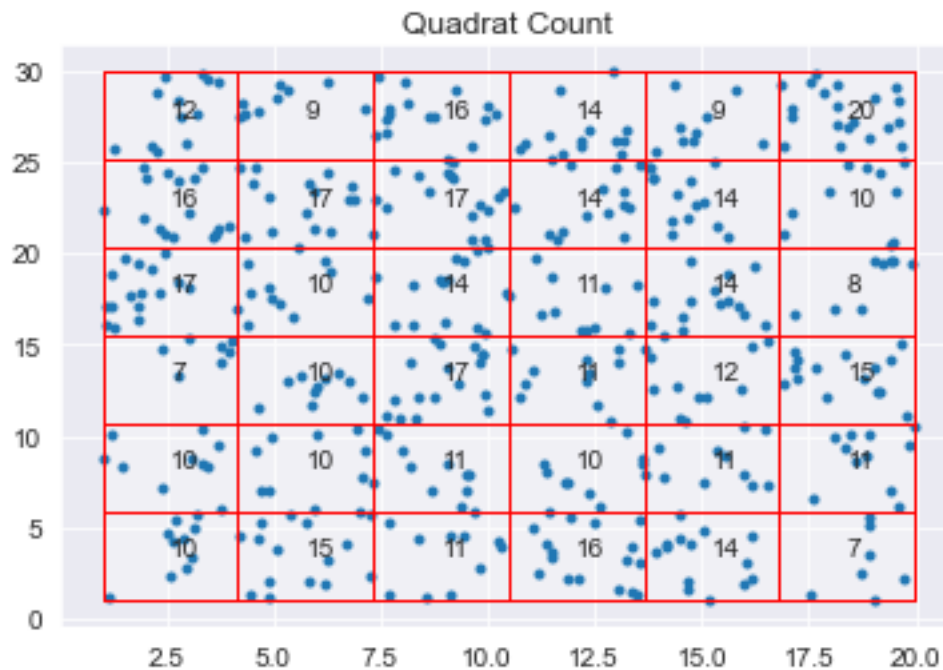


Figure 4-20: Quadrat Density of Workers on Site

The p-value is then calculated for the quadrat density to indicate spatial randomness or quadrat density. If the p-value is less than 0.05 then the workers are said to be clustered as they have higher densities in separate quadrats than would be expected for worker that are randomly distributed.

The p-value for this site, **0.750598**, is indicative of the spatial randomness and random densities of workers in the different quadrats on site. Nevertheless, quadrat density measures are highly sensitive to the choice of quadrat size, thus, further density analysis has to be performed for more accurate measures of densities.

b. Voronoi-based Analysis:

Voronoi analysis tends to be a visual indicator of workers' densities. The assumption made here is that the Voronoi cell around the workers represents the workers working space. Hence, as the Voronoi cells around each worker gets smaller, the working radius of the working gets smaller, indicating that the worker is surrounded by a larger number of workers. This in turn shows the areas on site of potentially high and low density of workers. For this site, the Voronoi analysis carried out yields the plot shown in Figure 4-21.

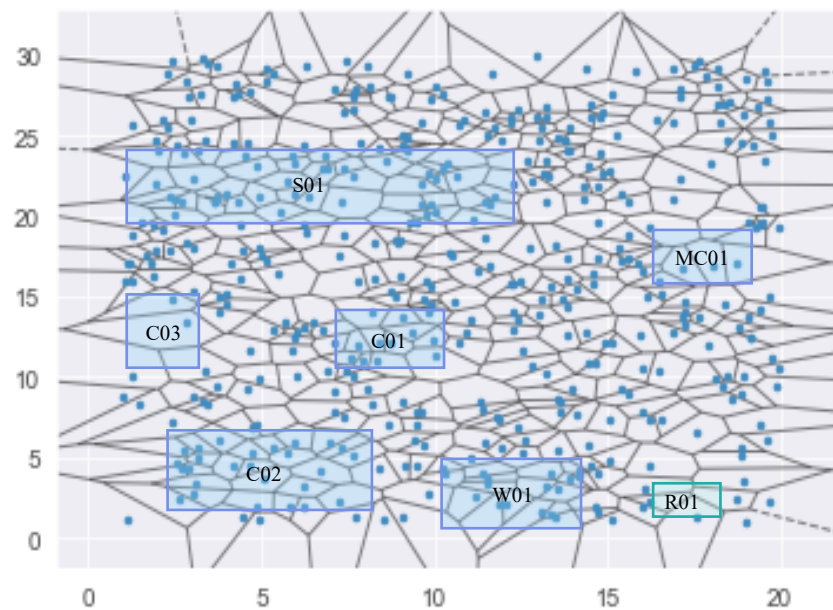


Figure 4-21: Voronoi Diagram of Workers on Site

In the figure above, the points represent the workers, and the lines represent the boundaries of the Voronoi cells around each worker. Voronoi diagrams are better representations of densities rather than a scatterplot of workers, however, are less desirable than heatmaps since they can be more difficult to visualize as the number of workers' coordinates increases.

c. Kernel Density Estimation – KDE:

Similar to the KDE contour plots produced for safety and quality, heatmaps are generated for workers' density to identify the zones of high productivity, medium productivity, and low productivity. These zones are identified assuming that the concentration of workers is indicative of the productivity of the zone on site. The intensity of the contour represents the density of the workers, the more intense the color of the contour, the higher the density of the workers in that specific location, the higher the productivity is anticipated to be. The heatmap for the workers density on the given site is shown in Figure 4-22.

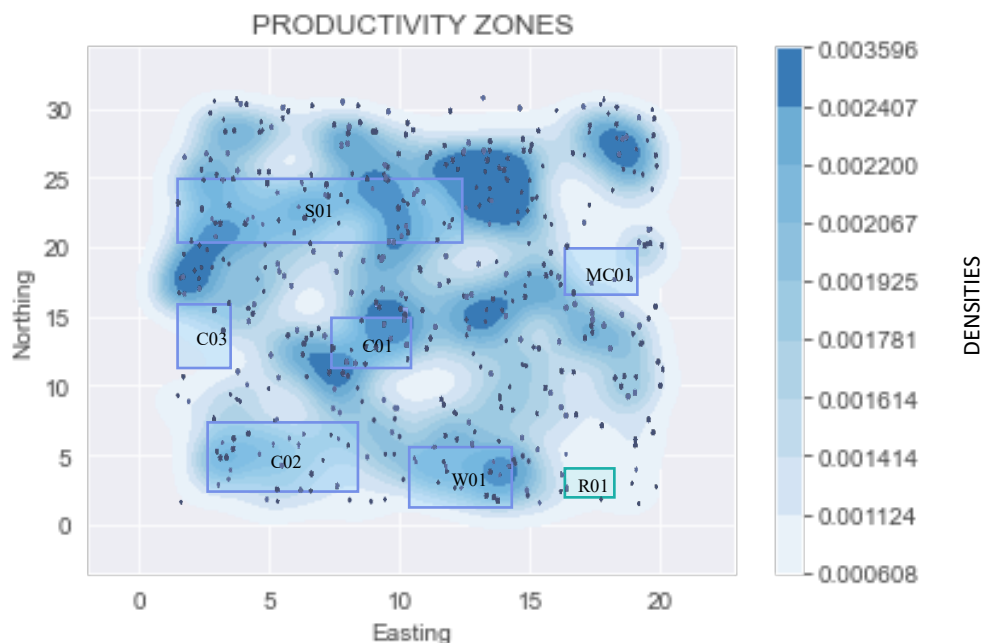


Figure 4-22: Productivity Zones - Heatmap of Workers' Density on Site

The color bar of the heatmap shows the probability densities of the workers in the spatial 2-D plane. The zones of high intensity contours are identified to be high productivity zones. The aforementioned assumption shall be verified later when performing Second Order Correlative Analysis.

4.2.1.3.4 Workers' Time Distribution

The last analysis done under the developed framework, is the workers' time distribution. Knowing the location of the worker, the zone in which the worker is located can be identified. Thus, knowing the category of the zone in which the workers are located and by means of the logic flow-chart shown in Figure 3-26, the percentage of time spent by the workers in each zone is calculated. These percentages are used to give an indication of the overall productivity on site. Usually higher percentages of working time, and lower percentages of traveling time are preferable on a construction site, because this means that there is less time wasted on site. The workers' time distribution for the generated site is shown in Figure 4-23.

B	0.816697
S01	0.076225
C02	0.047187
W01	0.027223
C01	0.018149
MC01	0.007260
C03	0.003630
R01	0.003630
Name: Area_Code, dtype: float64	
TA	0.816697
WA	0.179673
RA	0.003630

Figure 4-23: Workers' Time Distribution on Site

The results from the distribution firstly shows the percentage of time spent by workers in each individual zone, then the percentages are calculated for the category if each zone, whether it's a working zone, a traveling zone, or a resting zone. For this site, the workers spend around **81.7% of their time traveling, 18% working, and 0.3% resting**. This is not the most desirable for a working site, however, this is due to the random generation of workers data.

4.3 Stage 3 – Second Order Data Analysis (*SOCDA*)

SOCDA is performed with the purpose of obtaining further analysis about the site using the outputs from the *FODA*. The further analysis aims to depict any correlations or relationships between the different site parameters. The results from the *FODA* are compared against each other according to the matrix shown earlier in Table 4-1. The outputs from the *SOCDA* could be split into 5 main correlation categories. These categories are:

1. Correlation between Safety Performance and Productivity.
2. Correlation between Quality Performance and Productivity.
3. Correlation between Site Layout and Productivity.
4. Correlation between Project Schedule and Productivity.
5. Correlation between Cost Expenditure and Productivity.

The purpose of this section is not to provide an empirical method of detecting correlation; rather, it is to demonstrate the basic visual or comparative approach that could be used rationally to detect the existence of relationships between the site parameters.

4.3.1 Correlation between Safety Performance and Productivity

By comparing the outputs from Safety Data Analysis and Productivity Data Analysis, it could be identified whether there exists a relationship between the workers' spatial behavior and the site's safety performance. The relationship can be detected by comparing the following *FODA* outputs:

1. Workers' Densities vs Site Safety Zones
2. Workers' Central Tendencies vs Site Safety Zones
3. Workers' Clustering vs Site Safety Performance
4. Workers' Time Distribution vs Site Safety Performance
5. Workers' Activity vs Site Safety Behavior

The results from each of the comparisons will either confirm or refute the hypothesis of the presence of a connection between the parameters.

4.3.1.1 Workers' Densities vs Safety Zones

Comparing the locations of workers' densities against the safety zones on site, could provide an indication of whether workers' densities and safety performance are in anyway interrelated. If the location of high or low densities of workers on a site are visually perceived to be at the same locations of any of the safety risk zones, then the hypothesis of an existing relationship is confirmed.

By examining the productivity zones heatmap, voronoi diagram, and safety zones heatmap, as shown in Figure 4-24, it could be inferred that locations of high and low worker densities do not coincide with the locations of safety risk zones. Thus, there seems to be no direct relationship between the crowding of workers and the safety incidents occurring on site. Hence, the null hypothesis is rejected.

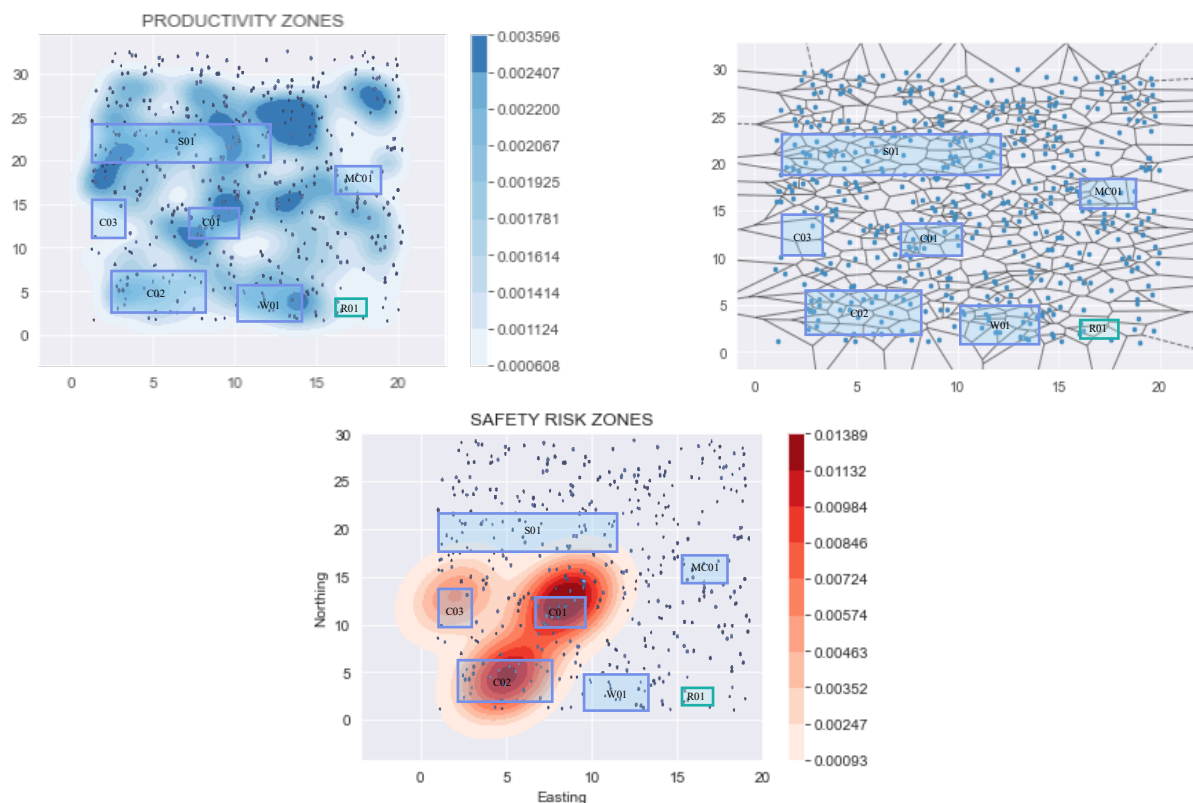


Figure 4-24: Workers' Density vs Site Safety Zones

4.3.1.2 Workers' Central Tendencies vs Safety Zones

Moreover, the central tendencies of the workers could be studied against the site safety zones to further test the null hypothesis. If the central measures of the workers match the with the safety zones on site, that means that the workers' spatial behavior on site influences the site's safety performance.

The central mean and median could be analyzed in reference to the locations of safety zones on site as displayed in Figure 4-25.

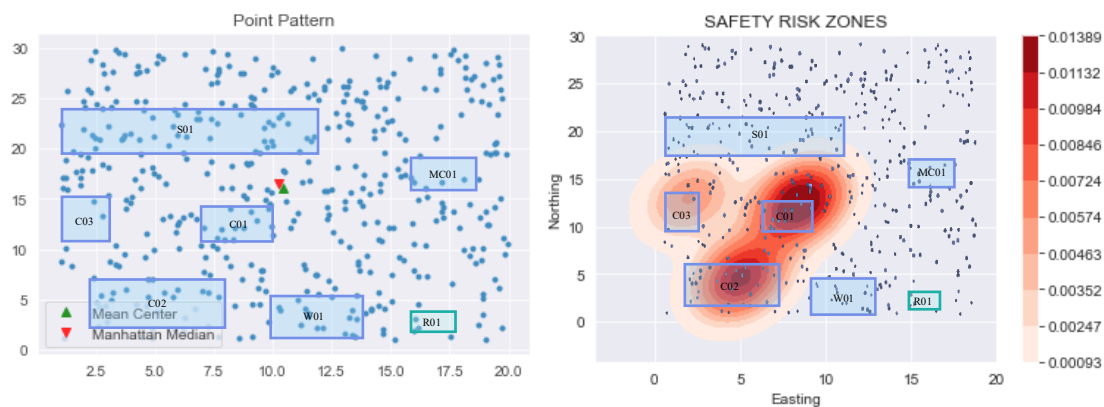


Figure 4-25: Workers' Central Tendencies vs Site Safety Zones

Generally, as noted from the figure, the location around which the workers were centralized over the working time period does not coincide with the location of safety high-risk or low-risk zones on site. This indicates that the centralization of workers, from the randomly generated coordinates, does not affect the site's safety performance. Hence, the null hypothesis is rejected for the random data generated.

4.3.1.3 Workers' Clustering vs Site Safety Performance

Knowing whether the workers are clustered or dispersed on site, could be used to detect if this influences the site's safety performance. This could be done by comparing the p-value, used to indicate workers clustering, against the mean and the median of the safety incidents occurring on site. If a site has workers that are clustered and has a high mean or median of safety incidents, then this could potentially mean that the more the workers are clustered, the more safety incidents occur on site as a result of overcrowding. Otherwise, if the workers are dispersed and the mean/median of safety incidents is high, this could mean that the dispersion of workers allows for more incidents to take place as there is less supervision or monitoring in the area.

4.3.1.4 Workers' Time Distribution vs Site Safety Performance

In addition to using workers' clustering, the distribution of the workers' time on site could be adopted for the comparative analysis. The percentage of time a worker spends working, traveling, or resting, could be compared with the safety performance on site to detect whether a relationship between the parameters exists or not. Analyzing the percentages against the mean/median of safety incidents, may perhaps signify the existence of a link. For example, if workers on a site spend most of their time working, and that site has a high mean/median of safety incidents, then this would indicate that the incidents tend to occur mainly in areas where activities are on-going. However, if the same is true for workers spending their time traveling, then this would indicate that there is no relationship between the working activities on site and the occurrence of safety incidents on site.

4.3.1.5 Working Activity vs Site Safety Behavior

Finally, by knowing the workers' spatial data, and the activity which the worker performs, a relationship between the working activity and the site's safety behavior could be identified. The moving average of the safety incidents is studied in conjunction with the type of activity on-going on site. The type of activity is indicated by the coordinates of the workers involved in the on-going activity, i.e., the coordinates of concrete pouring labor on site imply that the on-going activity is concrete pouring. Comparing the working activity with the moving average. could suggest a connection between the type of activity and the average of safety incidents that take place on site. The existence of a connection could potentially be used to recognize which activities on site are the most hazardous.

The comparison is done by examining the workers' coordinate according to their working activity on a given day with the moving average calculated for that day as shown in Figure 4-26.

The figure indicates that on the given day, 22nd of December 2020, the moving average increased due to the high number of incidents occurring on that day. On the same day, the observed workers' coordinates do not indicate the dominance of a specific activity. Accordingly, for the given data it cannot be concluded whether there is a relationship between the on-going activities and the safety site behavior.



Figure 4-26: Working Activity vs Site Safety Behavior

4.3.2 Correlation between Quality Performance and Productivity

It becomes possible to determine whether there is a relationship between workers' spatial behavior and the site's quality performance by comparing the results of Quality Data Analysis and Productivity Data Analysis. Comparing the following FODA outputs, the relationship can be discovered:

1. Workers' Densities vs Site Quality Zones
2. Workers' Central Tendencies vs Site Quality Zones
3. Workers' Clustering vs Site Quality Performance
4. Workers' Time Distribution vs Site Quality Performance
5. Workers' Activity vs Site Quality Behavior

The outcomes of each comparison will either confirm or refute the hypothesis of a connection between the parameters.

4.3.2.1 Workers' Densities vs Site Quality Zones

When the locations of workers' densities are compared to the quality zones on site, it may be possible to determine whether workers' densities and quality performance are in any way related. If the locations of high or low worker densities on a site are visually perceived to be in the same locations as any of the quality zones, the hypothesis of an existing relationship is confirmed.

By examining the productivity zones heatmap, voronoi diagram, and quality zones heatmap, as shown in Figure 4-27, it is possible to conclude that high and low worker densities do not coincide with quality zones. As a result, there appears to be no direct relationship between worker crowding and on-site inspection requests. As a result, the null hypothesis is rejected.

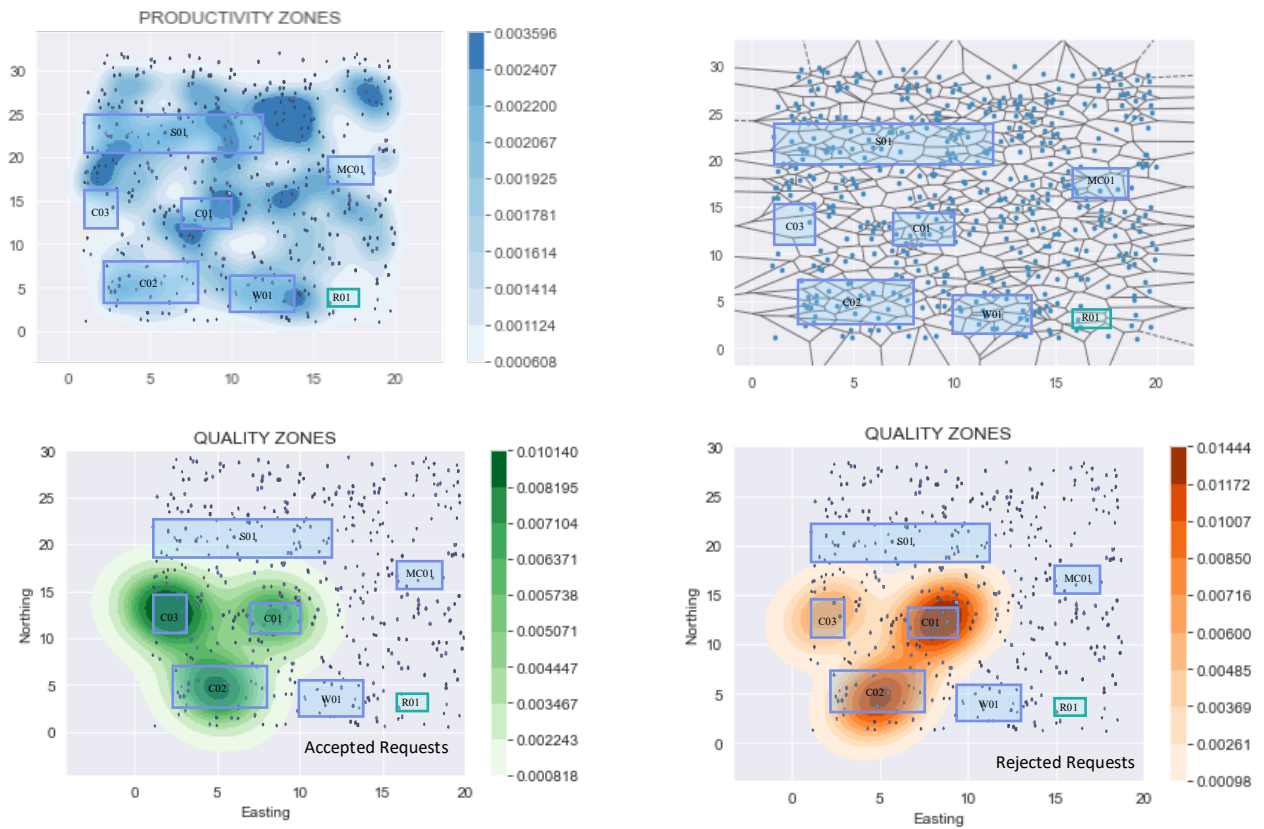


Figure 4-27: Workers' Density vs Site Quality Zones

4.3.2.2 Workers' Central Tendencies vs Site Quality Zones

Furthermore, the workers' central tendency might be compared to the site quality zones to further test the null hypothesis. If the workers' central measures correspond to the site's quality zones, this indicates that the workers' spatial behavior might have an impact on the site's quality performance.

The central mean and median could be analyzed in reference to the locations of quality zones on site as displayed in Figure 4-28.

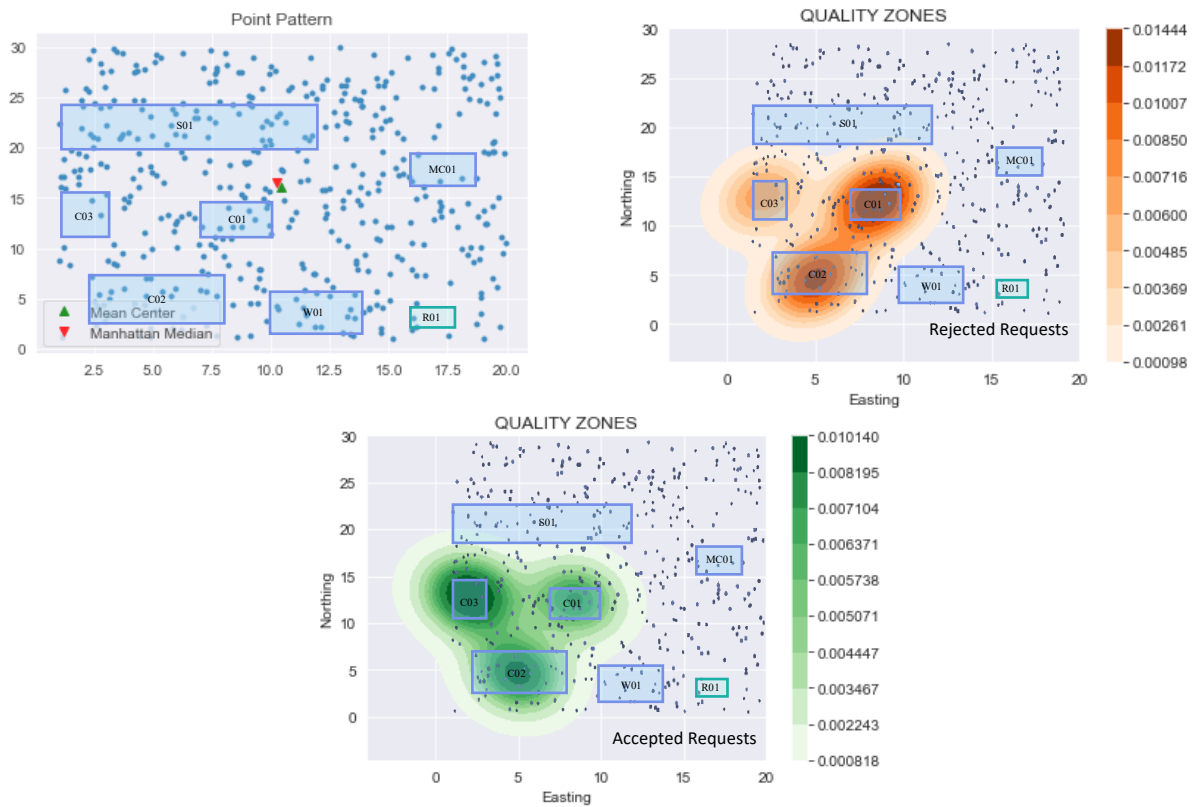


Figure 4-28: Workers' Central Tendencies vs Site Quality Zones

The position around which the workers were centralized over the working time period did not correlate with the location of high-quality or low-quality zones on site, as shown in the figure. This demonstrates that the site's quality performance is unaffected by worker centralization based on the randomly generated coordinates. As a result, the null hypothesis for the generated random data is rejected.

4.3.2.3 Workers' Clustering vs Site Quality Performance

Knowing whether workers are clustered or dispersed on the job site could be utilized to determine whether this has an impact on the site's quality performance. This might be accomplished by comparing the p-value, which is used to identify workers clustering, to the mean and median of the site's inspection requests. If a site has clustered workers and a high mean or median of rejected inspection requests, it's possible that the more clustered the workers are, the lower the quality of work executed on site as a result of overcrowding. If, on the other hand, the workers are distributed and the mean/median of rejected inspection requests is high, this could indicate that the workers' dispersion

allows for poor execution of work since there is less supervision or monitoring in the region.

4.3.2.4 Workers' Time Distribution vs Site Quality Performance

For the comparative study, in addition to worker clustering, the distribution of the workers' time on site could be used. To determine whether there is a relationship between the parameters, the percentage of time a worker spends working, travelling, or sleeping could be compared to the quality performance on site. Comparing the percentages to the mean/median of inspection requests could indicate the presence of a link. For example, if site personnel spend most of their time working and the site has a high mean/median of rejected inspection requests, this indicates that workmanship of personnel in the working areas is unsatisfactory. However, if the same is true for workers who spend their time commuting, this would show that there is a link between the time workers spend working and the likelihood of inspection requests being accepted or rejected.

4.3.2.5 Working Activity vs Site Quality Behavior

Finally, a relationship between the working activity and the site's quality behavior could be identified by knowing the workers' spatial data and the activity that the worker performs. The moving average of inspection requests is examined in conjunction with the type of activity taking place on site. The coordinates of the workers involved in the ongoing activity indicate the type of activity; for example, the coordinates of concrete pouring labor on site imply that the ongoing activity is concrete pouring. Comparing the working activity to the moving average could indicate a link between the type of activity and the average number of accepted or rejected inspection requests on site. The presence of a link could potentially be used to identify which activities on site are more onerous to execute and might require frequent rework.

The comparison is made by comparing the coordinates of the workers based on their working activity on a given day to the moving average calculated for that day, as shown in Figure 4-29.



Figure 4-29: Working Activity vs Site Quality Behavior

The figure shows that the moving averages increased on the given day, December 22nd, 2020, due to the high number of accepted and rejected that occurred on that day. The observed workers' coordinates on the same day do not indicate the dominance of a specific activity. As a result, it is not possible to conclude whether there is a relationship between ongoing activities and safety quality behavior based on the available data.

4.3.3 Other Site Parameters and Productivity

Finally, by examining the workers' spatial temporal behavior, other site parameters could be monitored or evaluated. These parameters are:

1. Efficiency of construction site layout
2. Construction progress variance
3. Value of money for site cost expenditure

The following outputs are used from the FODA to examine the obtain parameters:

1. Workers' Spatial Temporal Locations
2. Workers' Central Tendency
3. Workers' Density
4. Workers' Time Distribution

4.3.3.1 Efficiency of Construction Site Layout

According to the matrix in Table 4-1, by studying the workers' central tendencies, densities, and time distribution, the efficiency of the site layout could be determined.

a. Workers' Density:

The locations of worker densities may indicate the efficiency of site layouts. The closer the high-density worker areas are, the more effective the site layout. This means that more workers are present in the working areas and that workers are not wasting time commuting between the various site facilities.

b. Workers' Central Tendency:

Also, determining the central tendencies of workers may help in assessing the efficiency of the site layout. For an efficient site layout, the central mean and median of the workers should be located closer to the working areas rather than travelling. As a result, if workers tend to be less centralized within working areas, this could indicate that the different working areas are too distant, i.e., workers coordinates could be more spread out due to the inefficient layout of the various site facilities. This phenomenon could be confirmed further by examining the workers' time distribution.

For sites where the construction areas themselves are located far from each other, in that case the workers in the different areas might be identified as separate clusters. Examining the central mean or median of the separate clusters may indicate the efficiency of the site layout, since the central tendencies of each cluster are expected to be located near the working area of each cluster.

Also, the direction of the standard deviational ellipse might give an indication of the layout's efficiency. The larger the directional axis of the ellipse, the longer the paths the workers might have to take to reach the different site facilities, the more inefficient the site layout becomes, and vice versa.

c. Workers' Time Distribution:

Finally, the time distribution of the workers is the definitive indicator of the site's layout effectiveness. The greater the percentages of commuting time, the more time workers waste traveling from and to various working areas and site facilities. Lower percentages of traveling time would be calculated if the site layout was more efficient, indicating that less time is wasted in commuting.

4.3.3.2 Construction Progress Variance

The spatial temporal locations, central tendency, density, and time distribution of workers could all be deployed as measures or indicators of the construction progress on site at a given point in time.

a. Workers' Spatial Temporal Locations:

Being able to aggregate the workers' spatial location by the activity in which the worker is involved, allows for the identification of the on-going activity during a specific period. Therefore, identifying the locations of formwork and scaffolding labor within a site construction area, indicates that the formwork activity is on-going. The on-going activity could then be compared against the planned activity during that period according to baseline schedule of the project. If the planned activity is the same as the on-going activity in the area, then the area is on schedule. However, if the planned

activity is a succeeding or preceding activity to the on-going activity, then this means that the area is behind schedule or ahead of schedule respectively.

b. Workers' Density:

The densities of workers can also be used to indicate progress variances. Given high worker densities, these are assumed to be areas of high productivity or activity on site at a specific point in time. By comparing the locations of the densities to the locations of the areas that, according to the project's baseline schedule, should have the most activity, the project's progress deviation from the baseline can be identified. If the workers' high densities are in areas where work activity is not planned to be on-going at the time, then the progress of the project has deviated from the project's original baseline.

4.3.3.3 Value of Money for Site Cost Expenditure

Workers' spatial temporal locations, density, and time distribution could all be used as measures or indicators of the value of money for cost expenditures on site at any given point in time.

a. Workers' Spatial Temporal Locations:

The ability to identify the on-going activity by locating the workers involved in the execution of the activity allows for assessment of the cost of the activities on site. When the costs spent in a construction area over a specific period are compared to the identified on-going activities during the same period, the money spent on the activities can be quantified. These costs can then be compared to the activities' budgeted costs to determine whether the activities are over or under budget.

b. Workers' Density:

Worker densities can also be used to indicate value for money of the personnel costs spent on site. Given the high/low worker density, these are assumed to be areas of high/low activity or productivity on-site at any given time. The costs incurred in the areas could then be compared to locations with high and low densities to determine the value of money. If areas of high workers' densities have higher costs, then this could

indicate that the value of money spent is realized since the workers are active in these areas. This could be also compared against the quality zones on site for further verification. However, if areas of high costs have low densities of workers, this might mean that there are unnecessary costs being spent on the area that could be reduced, or that the costs of the personnel in the area are over-estimated for the productivity.

c. Workers' Time Distribution:

Finally, the workers' time distribution is the ultimate indicator of the value of money spent on a construction site. The greater the percentages of commuting or resting time, the more time workers waste travelling or resting and the fewer working hours they spend on site. As a result, most indirect costs incurred on-site are misspent as workers spend more time travelling and less time working.

4.4 Summary of Findings

This section aims to demonstrate the feasibility of implementing the framework for the analysis of data collected from construction sites. After the first order data analysis has been performed, the application of the second order correlative data analysis becomes achievable. Where the SOCDA is used to highlight the potential of deploying outputs from the FODA to determine the existence of relationships between spatial behavior of workers and site performance in terms of safety, quality, productivity, progress, site layout efficiency, and cost expenditure.

CHAPTER 5 – CASE STUDY

This chapter presents a case study project in Ain Al Sokhna, Egypt verifying the feasibility of implementing the framework on an actual construction site. The final outputs from implementing the framework are discussed to highlight the potential benefit of applying real-time monitoring technologies for data collection and site performance monitoring.

5.1 Project Information

The presented case study is for a commercial project in Ain Al Sokhna, Egypt. The project is called, New Capital Sportive Village and is part of the Egyptian Government's urban development plan for coastal cities. The contract value is around 4 billion Egyptian pounds and was planned to be completed in 4 years.

The main project is comprised of multiple sub-projects given the enormity of the building areas. These sub-projects are:

1. Hotel Area
2. Electronic Shooting Area
3. Cartouche Shooting Area
4. Admin Buildings
5. Manual Shooting Area
6. Commercial Area

Only one sub-project, the hotel area, was used for the implementation of the framework and the data was collected during the construction of the chosen sub-project.

5.2 Framework Implementation

The purpose of implementation is to demonstrate the feasibility of applying the framework, thus, the developed framework is applied for first order data analysis of real-time data collected only. The framework was implemented according to the steps shown in Figure 5-1. Firstly, a single site coordinate was obtained from the site's surveying team. The single coordinate was used to locate the site on a geographic

information system and obtain the remaining required site coordinates. Secondly, the pre-defined site zones were obtained from the construction team on site. Thirdly, multiple GPS tracking applications were tested to find the most efficient application to use. Finally, two participants downloaded and used the tracking application for a certain period of time, and their daily tracks were used for the analysis.

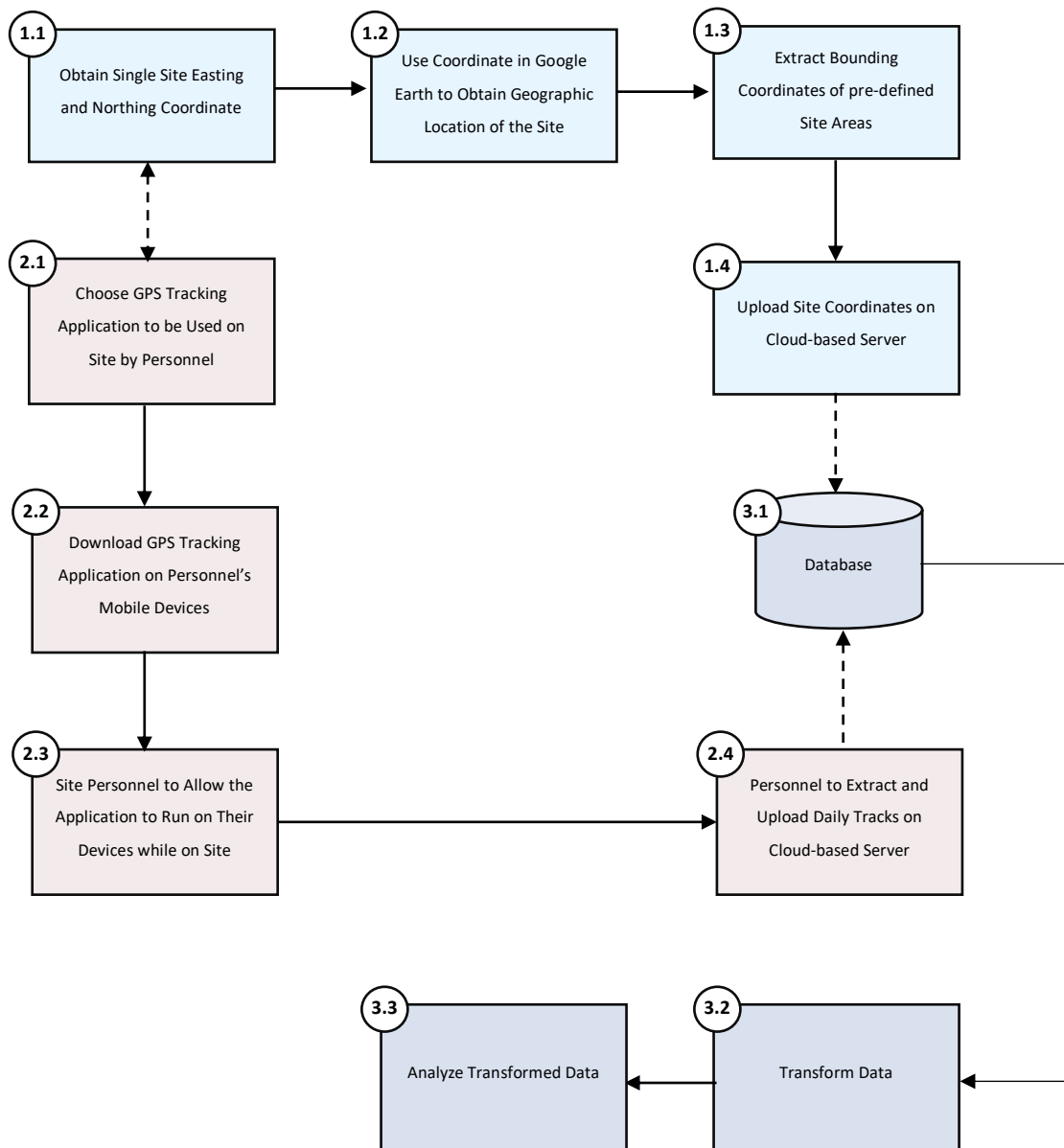


Figure 5-1: Framework Implementation Flowchart

5.2.1 Data Collection

Site geographic data and real-time workers' spatial data were collected from the construction site using the data collection methods explained under section 3.1 – *Stage 1 – Data Collection*.

5.2.1.1 Site Geographic Data

Using the single coordinate obtained from the site team, 29.90080957 and 31.68311107, Google Earth was used to locate the site geographically. The coordinate was searched for and the location was obtained as shown in Figure 5-2.



Figure 5-2: View of Construction Site obtained from Google Earth

After obtaining the sites' geographical location, the specific coordinates of the site boundaries as well as the bounding coordinates of the different site areas were extracted in the format shown in Figure 5-3. The category of each area was predefined by the construction team on site.

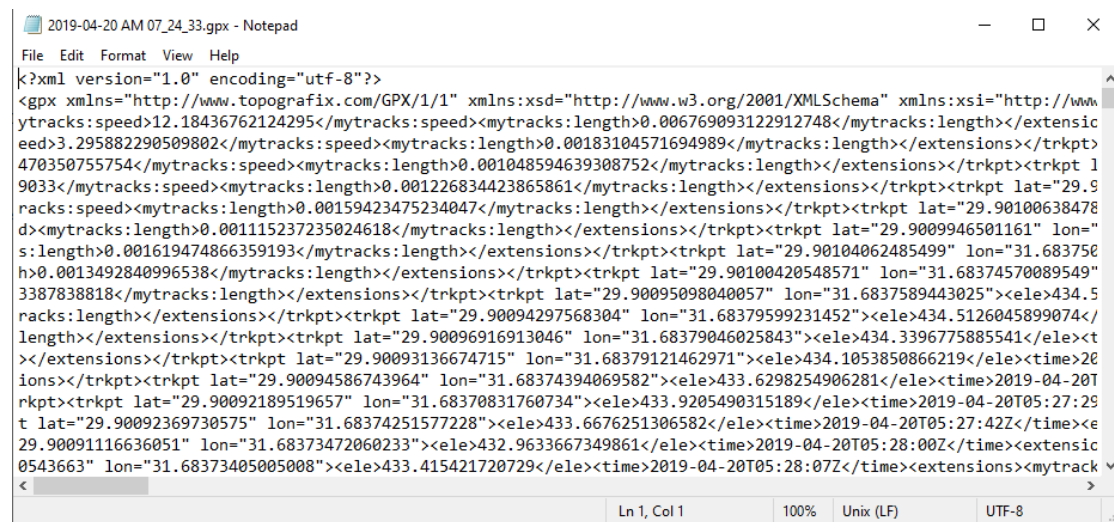
ns1:name3	ns1:styleUrl4	ns1:tessellate	ns1:coordinates
Untitled Polygon	#m_ylw-pushpin		1 31.68147503337467,29.89952583971098,0 31.68161271112465,29.89952584213956,0 31.681606

Figure 5-3: Extracted Bounding Coordinates of Site Areas

5.2.1.2 Real-time Geospatial Data

For the real-time workers data, the myTracks app, previously discussed in section 3.1.3.1 - *Smart Phone GPS Tracking Application of Choice & Its Implementation*, was downloaded on the work phones of 1 Senior and 1 Junior Site engineer supervising the construction of the hotel area.

The engineers would start allowing the app to track their locations once they enter the site, and let the application run for the entire working day. Once their working day is complete, the engineers would extract their tracks in gpx formats and upload it on the SharePoint created for the framework. The extracted gpx files are shown in Figure 5-4.



```
2019-04-20 AM 07_24_33.gpx - Notepad
File Edit Format View Help
<?xml version="1.0" encoding="utf-8"?>
<gpx xmlns="http://www.topografix.com/GPX/1/1" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www
mytracks:speed>12.18436762124295</mytracks:speed><mytracks:length>0.006769093122912748</mytracks:length></extensic
eed>3.295882290509802</mytracks:speed><mytracks:length>0.00183104571694989</mytracks:length></extensions></trkpt>
470350755754</mytracks:speed><mytracks:length>0.001048594639308752</mytracks:length></extensions></trkpt><trkpt 1
9033</mytracks:speed><mytracks:length>0.001226834423865861</mytracks:length></extensions></trkpt><trkpt lat="29.9
racks:speed><mytracks:length>0.00159423475234047</mytracks:length></extensions></trkpt><trkpt lat="29.90100638478
d><mytracks:length>0.001115237235024618</mytracks:length></extensions></trkpt><trkpt lat="29.9009946501161" lon="
s:length>0.001619474866359193</mytracks:length></extensions></trkpt><trkpt lat="29.90104062485499" lon="31.683750
h>0.0013492840996538</mytracks:length></extensions></trkpt><trkpt lat="29.90100420548571" lon="31.68374570089549"
3387838818</mytracks:length></extensions></trkpt><trkpt lat="29.90095098040057" lon="31.6837589443025"><ele>434.5
racks:length></extensions></trkpt><trkpt lat="29.90094297568304" lon="31.68379599231452"><ele>434.5126045899074</
length></extensions></trkpt><trkpt lat="29.90096916913046" lon="31.68379046025843"><ele>434.3396775885541</ele><t
></extensions></trkpt><trkpt lat="29.90093136674715" lon="31.68379121462971"><ele>434.1053850866219</ele><time>20
ions></trkpt><trkpt lat="29.90094586743964" lon="31.68374394069582"><ele>433.6298254906281</ele><time>2019-04-20T
rkpt><trkpt lat="29.90092189519657" lon="31.68370831760734"><ele>433.9205490315189</ele><time>2019-04-20T05:27:29
t lat="29.90092369730575" lon="31.68374251577228"><ele>433.6676251306582</ele><time>2019-04-20T05:27:42Z</time><e
29.90091116636051" lon="31.68373472060233"><ele>432.9633667349861</ele><time>2019-04-20T05:28:00Z</time><extensic
0543663" lon="31.68373405005008"><ele>433.415421720729</ele><time>2019-04-20T05:28:07Z</time><extensions><mytrack
<
Ln 1, Col 1    100%    Unix (LF)    UTF-8
```

Figure 5-4: gpx File Extracted for a Single Engineer on Site

For verification of the workers' location, the gpx files were uploaded on Google Earth and the daily tracks were viewed as shown in Figure 5-5.



Figure 5-5: The Engineer's Track on Site Viewed Using Google Earth

The figure shows that the workers are correctly located on site, thus providing further verification of the accuracy of the app used for tracking and the definition of site boundaries.

5.2.2 Data Preparation and Cleaning

The data collected was not processable directly by the analysis algorithms, hence, the data needed to be transformed into a unified format capable of being handled by the algorithm.

5.2.2.1 Site Geographic Data Preparation

The site geographic data obtained was transformed using manual grouping where, the coordinates for each area were grouped to create the bounding polygon of the zone. Each set of coordinates was referred to the unique area of the code as well as the predefined area's category as shown in Table 5-1.

Table 5-1: Transformed Bounding Coordinates of Site Areas

Area_Code	Ref_No.	Easting	Northing	Elevation	Category
C01	1	31.673178	29.900439	400	WA
C01	2	31.673178	29.899791	400	WA
C01	3	31.674391	29.899791	400	WA

C01	4	31.674391	29.899318	400	WA
C01	5	31.678397	29.899318	400	WA
C01	6	31.678397	29.899785	400	WA
C01	7	31.679774	29.899785	400	WA
C01	8	31.679774	29.900595	400	WA
C01	9	31.673178	29.900595	400	WA
C01	10	31.673178	29.900439	400	WA
B	11	31.671130	29.890400	400	TA
B	12	31.686560	29.890400	400	TA
B	13	31.686560	29.901790	400	TA
B	14	31.671130	29.901790	400	TA
C02	15	31.681475	29.899526	400	WA
C02	16	31.681613	29.899526	400	WA
C02	17	31.681607	29.899288	400	WA
C02	18	31.682437	29.899290	400	WA
C02	19	31.682437	29.900086	400	WA
C02	20	31.682624	29.900087	400	WA
C02	21	31.682620	29.899555	400	WA
C02	22	31.682765	29.899555	400	WA
C02	23	31.682761	29.900031	400	WA
C02	24	31.683069	29.900031	400	WA
C02	25	31.684407	29.899501	400	WA
C02	26	31.684430	29.899395	400	WA
C02	27	31.684535	29.899388	400	WA
C02	28	31.684527	29.899584	400	WA
C02	29	31.683824	29.899869	400	WA
C02	30	31.683996	29.900195	400	WA
C02	31	31.683515	29.900394	400	WA
C02	32	31.683507	29.900862	400	WA
C02	33	31.683824	29.900864	400	WA
C02	34	31.683914	29.900868	400	WA
C02	35	31.683930	29.901191	400	WA
C02	36	31.683945	29.901231	400	WA
C02	37	31.683872	29.901245	400	WA
C02	38	31.682761	29.901247	400	WA
C02	39	31.682752	29.900924	400	WA
C02	40	31.682743	29.900467	400	WA
C02	41	31.682273	29.900270	400	WA
C02	42	31.682274	29.900066	400	WA
C02	43	31.681477	29.900066	400	WA
C02	44	31.681475	29.899526	400	WA
C03	45	31.680988	29.897196	400	WA
C03	46	31.680660	29.897196	400	WA
C03	47	31.680081	29.897199	400	WA
C03	48	31.679120	29.897198	400	WA
C03	49	31.679139	29.896117	400	WA
C03	50	31.679384	29.896106	400	WA
C03	51	31.679406	29.895514	400	WA

C03	52	31.679689	29.895513	400	WA
C03	53	31.679690	29.895351	400	WA
C03	54	31.679386	29.895355	400	WA
C03	55	31.678972	29.895351	400	WA
C03	56	31.678971	29.894910	400	WA
C03	57	31.679700	29.894907	400	WA
C03	58	31.679716	29.894681	400	WA
C03	59	31.679580	29.894672	400	WA
C03	60	31.679590	29.894538	400	WA
C03	61	31.680100	29.894537	400	WA
C03	62	31.680080	29.895045	400	WA
C03	63	31.680511	29.895048	400	WA
C03	64	31.680521	29.894559	400	WA
C03	65	31.682019	29.894563	400	WA
C03	66	31.682012	29.898725	400	WA
C03	67	31.680951	29.898723	400	WA
C03	68	31.680988	29.897196	400	WA
MC01	69	31.670126	29.894809	400	WA
MC01	70	31.669737	29.894809	400	WA
MC01	72	31.669737	29.893524	400	WA
MC01	71	31.670126	29.893524	400	WA
W01	73	31.671023	29.894067	400	WA
W01	74	31.671023	29.893223	400	WA
W01	76	31.672880	29.893223	400	WA
W01	75	31.672880	29.894067	400	WA
S01	77	31.670117	29.893248	400	WA
S01	78	31.670117	29.890118	400	WA
S01	80	31.668752	29.890118	400	WA
S01	79	31.668752	29.893248	400	WA

For the site studied, the areas used in the framework are as shown in the table above.

These areas are:

1. Construction Area – C01: The cartouche building.
2. Construction Area – C02: The hotel building.
3. Construction Area – C03: The electronic shooting building
4. Main Storage Area – S01: The project's warehouse.
5. Main Caravan – MC01: The project's main caravan where all indirect manpower is located.
6. Workshop – W01: The project's main rebar carpentry workshop.

5.2.2.2 Real-time Geospatial Data Preparation

For the real-time data, a Python ® splitting technique was used to create a data frame of information from the gpx file. The data frame contained data regarding the employees and the employees' spatial temporal data as shown in Table 5-2.

Table 5-2: Sample of Engineers' Coordinates

Employee_ID	Employee_Type	Working_Activity	Easting	Northing	Elevation	Distance	Speed	Time	Date
E02	Senior Engineer	Supervision	31.682476	29.900542	425.208313	-	-	11:31:01	4/9/2019
E02	Senior Engineer	Supervision	31.682429	29.900498	414.087429	0.006606	23.781684	11:31:02	4/9/2019
E02	Senior Engineer	Supervision	31.682473	29.900567	410.104580	0.008699	15.658938	11:31:04	4/9/2019
E02	Senior Engineer	Supervision	31.682439	29.900602	406.599819	0.005138	18.496515	11:31:05	4/9/2019
E02	Senior Engineer	Supervision	31.682469	29.900517	411.010098	0.009861	35.499384	11:31:06	4/9/2019
E02	Senior Engineer	Supervision	31.682399	29.900527	402.509731	0.006767	24.363942	11:31:07	4/9/2019
E02	Senior Engineer	Supervision	31.682415	29.900589	408.956142	0.007073	12.731684	11:31:09	4/9/2019
E02	Senior Engineer	Supervision	31.682401	29.900623	408.867764	0.003971	14.301227	11:31:10	4/9/2019
E02	Senior Engineer	Supervision	31.682406	29.900639	408.324856	0.001831	3.291449	11:31:12	4/9/2019
E02	Senior Engineer	Supervision	31.682408	29.900648	407.693813	0.001091	3.926605	11:31:13	4/9/2019
E02	Senior Engineer	Supervision	31.682411	29.900660	407.048428	0.001277	4.598006	11:31:14	4/9/2019
E02	Senior Engineer	Supervision	31.682415	29.900671	407.076870	0.001376	4.976204	11:31:15	4/9/2019
E02	Senior Engineer	Supervision	31.682418	29.900687	407.222378	0.001743	6.274009	11:31:16	4/9/2019
E02	Senior Engineer	Supervision	31.682421	29.900704	407.066799	0.001884	6.801079	11:31:17	4/9/2019
E02	Senior Engineer	Supervision	31.682424	29.900721	407.036221	0.001957	7.053993	11:31:18	4/9/2019
E02	Senior Engineer	Supervision	31.682429	29.900740	406.927944	0.002216	7.978633	11:31:19	4/9/2019
E02	Senior Engineer	Supervision	31.682437	29.900755	406.661709	0.001809	6.511537	11:31:20	4/9/2019
E02	Senior Engineer	Supervision	31.682443	29.900768	406.629116	0.001560	5.615806	11:31:21	4/9/2019
E02	Senior Engineer	Supervision	31.682450	29.900780	406.405056	0.001510	5.446871	11:31:22	4/9/2019
E02	Senior Engineer	Supervision	31.682458	29.900791	406.260708	0.001452	5.228031	11:31:23	4/9/2019
E02	Senior Engineer	Supervision	31.682466	29.900803	406.370571	0.001433	5.160570	11:31:24	4/9/2019
E02	Senior Engineer	Supervision	31.682472	29.900815	406.360195	0.001543	5.574052	11:31:25	4/9/2019
E02	Senior Engineer	Supervision	31.682478	29.900826	406.036098	0.001329	4.760264	11:31:26	4/9/2019
E02	Senior Engineer	Supervision	31.682482	29.900836	405.893276	0.001186	4.268669	11:31:27	4/9/2019

Each engineer was given a unique ID for identification purposes and was assigned a working activity, which in their case is of type supervision.

5.2.3 First Order Data Analysis

After the data was transformed, the first order data analysis of the real-time data collected was performed. The analysis algorithms were applied using the Python ® programming language.

5.2.3.1 Productivity Data Analysis

Under this section, the real-time data collected from the site engineers, over a 10-day working period, $t = 10$ days, is used for implementing the first order analysis stage of the framework.

5.2.3.1.1 Visualization and Noise Removal of Workers' Location

a. Visualization:

Nevertheless, prior to the analysis, visualization of the workers' spatial locations and tracks was performed with the purpose of better understanding of the results from the analysis. By visualizing the workers' locations, the results can be verified against patterns and inferences made from examining the spatial behavior of workers over a period of time. The collected workers' data was visualized as shown in Figure 5-6, for the entire time period, and Figure 5-7, on the 4th of October 2019.

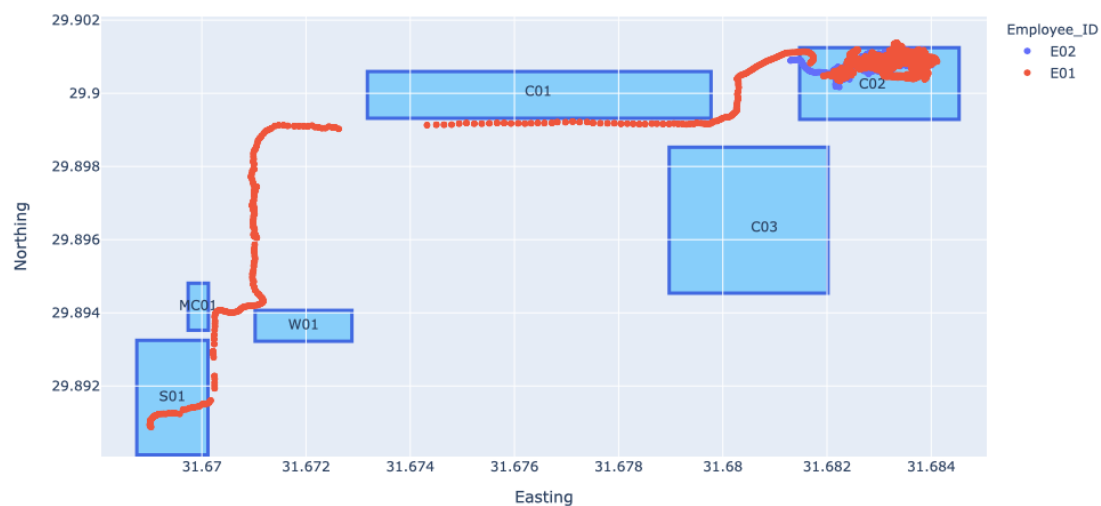


Figure 5-6: Engineers' Coordinates on Site over the 10-day Working Period

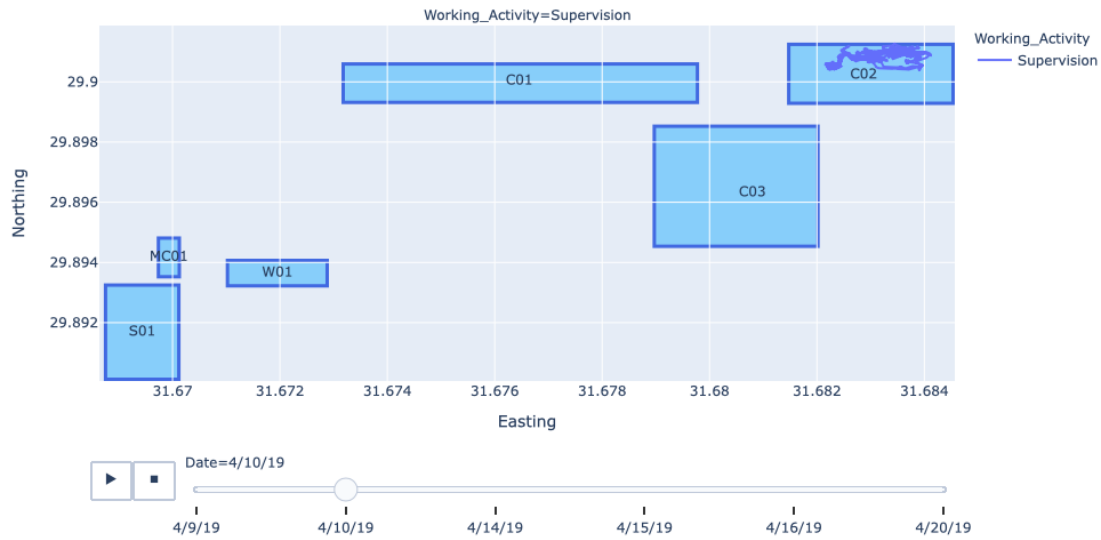


Figure 5-7: Engineers' Tracks on Site over the 10-day Working Period

As anticipated, since both workers are responsible for the supervision of construction works in the hotel area, most of their coordinates lie within the boundaries of area C02.

b. Noise Removal:

Even though the data could be analyzed and visualized without filtering, the process becomes inefficient and time consuming, and the results are more affected by noise data points. Accordingly, to enhance the visualization and analysis processes, the noise data points are removed using the *Ramer-Douglas-Peucker (RDP)* algorithm. The algorithm works by decimating a polyline, made up of segments, to a similar polyline with fewer points. The purpose of the algorithm is to eliminate unnecessary data points without significantly affecting the results of the analysis carried out on the filtered data points. The removal of noise becomes more beneficial as the datapoints increase.

Thus, the algorithm is used to filter the coordinates of the workers for an improved analysis process. The visualization of the data should be improved as well. Figure 5-8 shows the scatterplots of the workers after the RDP algorithm was applied on the workers coordinates.

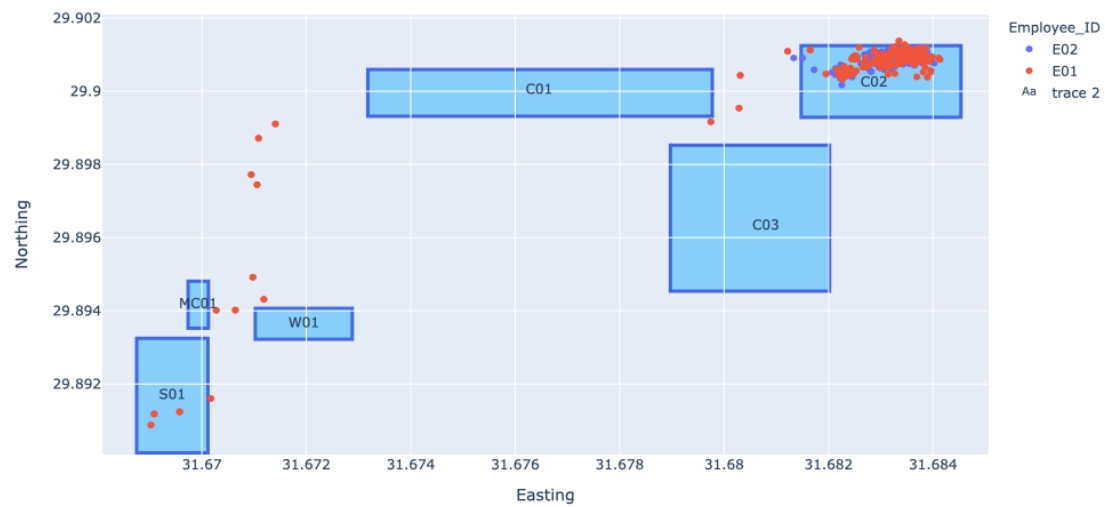


Figure 5-8: Engineers' Coordinates after RDP Algorithm Application

As seen the true coordinates of the workers are more visible after RDP analysis as the points were reduced from **23,869** coordinates to **792** coordinates, compressing the data by **96.7%**. Regardless, the final outputs of the analysis are not highly affected by the filtration of noise as will be seen later, while reducing the processing time by **90.4%**.

5.2.3.1.2 Central Tendency of Workers on Site

Testing the central tendency of both engineers on site, the central measures obtained are as follows:

c. Mean Center:

The coordinates of mean center calculated for the engineers are **(31.68311107, 29.90080957)**, and the identified mean center is as shown in Figure 5-9.

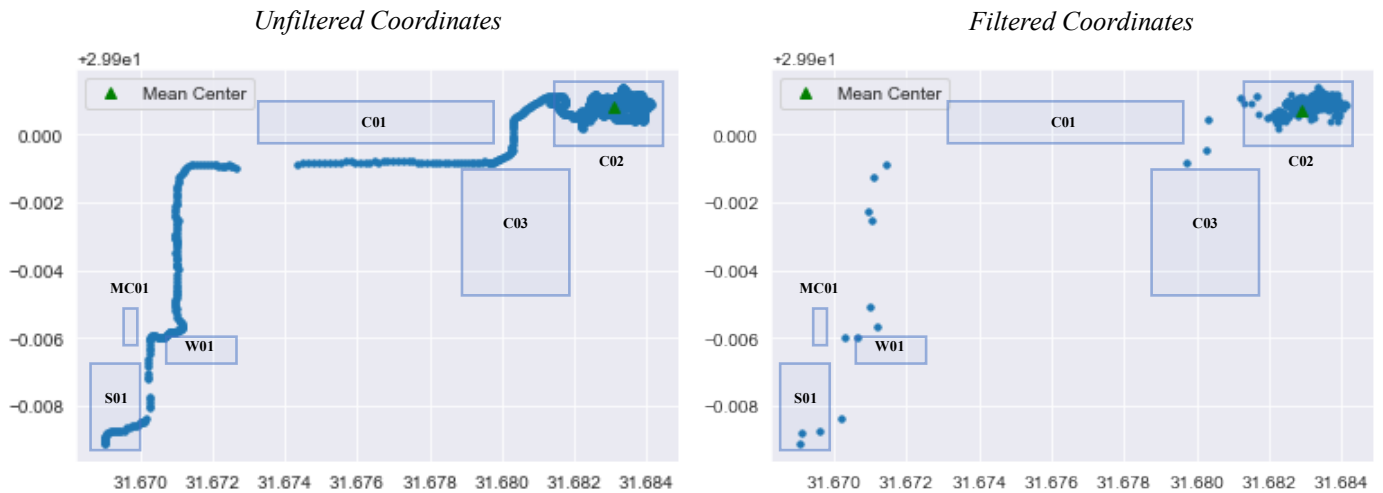


Figure 5-9: Spatial Mean Centers of Engineers

The location of the mean center, before RDP analysis, almost coincides with that of the center of C02 (31.682870, 29.900224), which indicates that the engineers were mostly located in C02 during the 10-day period. After the RDP algorithm is applied, the mean center was calculated to be (31.68290094, 29.90070515), verifying that the RDP did not cause any alterations to the results of the analysis.

d. Median Center:

As depicted from Figure 5-10, the median center does not coincide the mean center as on one day, an engineer spent most of his time outside of the working area, thus shifting the median center to (31.683546, 29.900947). The median center after RDP algorithm application became (31.683504, 29.900932).

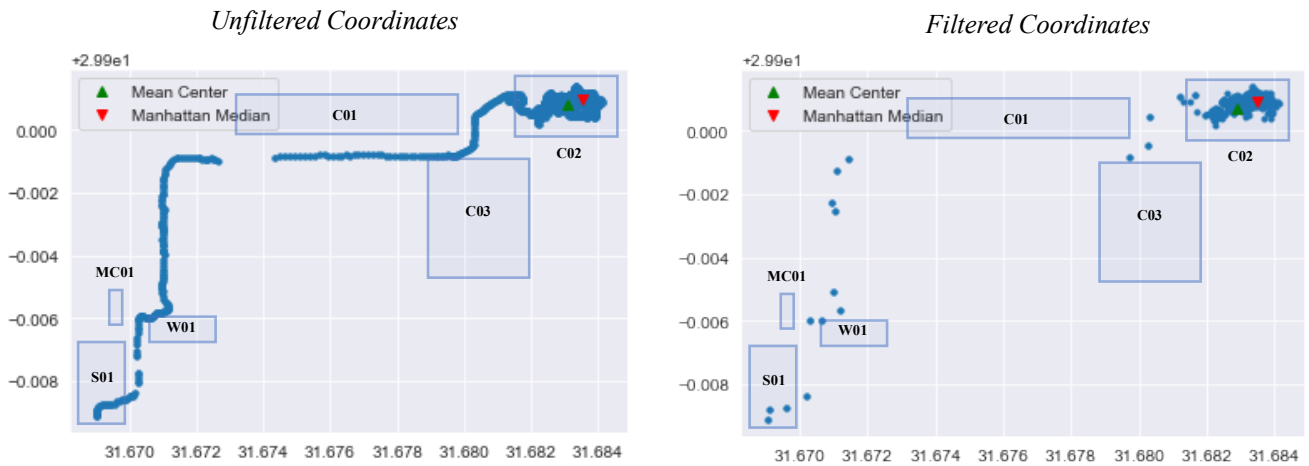


Figure 5-10: Median Centers of Engineers' Coordinates

However, the mean centers and median centers both were noticed to be near the geographic center of C02, verifying the engineers' spatial presence within the area.

e. Standard Distance:

The measured standard distance was calculated to be **0.001873 kilometers**, for unfiltered data, and **0.002578 kilometers** for filtered data. These were the radii for the standard circles encompassing most of the engineers' coordinates as shown in Figure 5-11.

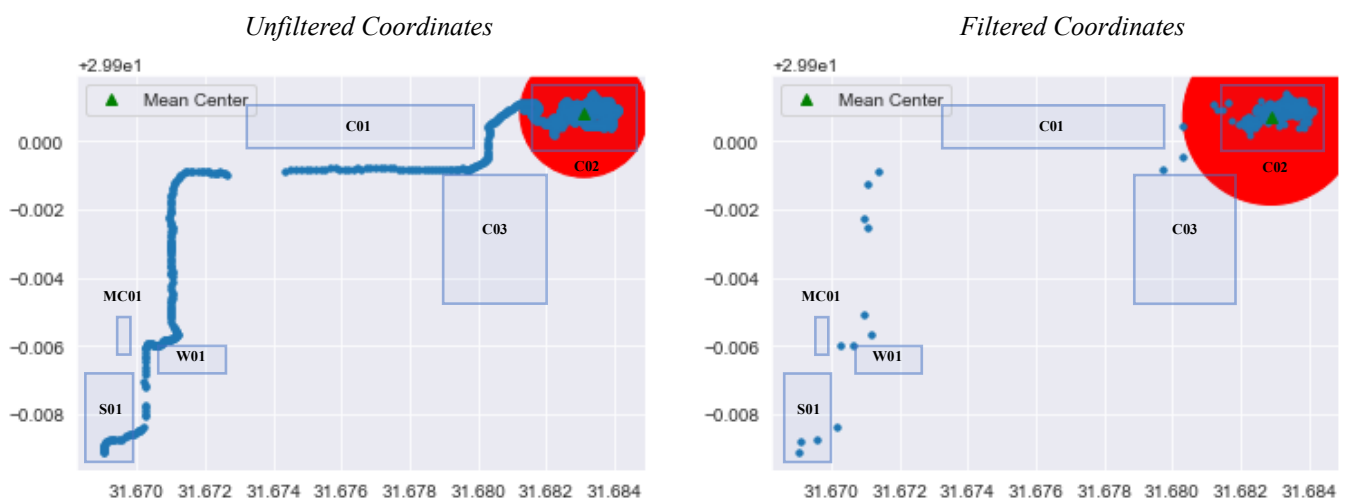


Figure 5-11: Standard Circles for Engineers' Coordinates

The standard circles are located mainly around the area of C02, again meaning that the engineers were mostly located in the working area.

f. Standard Deviational Ellipse:

Finally, the standard deviation ellipses shown in Figure 5-12 indicate the directional dispersion of the engineers. The ellipses are identical for unfiltered and filtered coordinates. As seen, the larger axes of the both ellipses are in the west-south and east-north direction. The detected directional dispersion is justified by the site layout since the main storage area and workshop are located on the far west-south corner of the site, and the C02 is located on the far east-north corner. Accordingly, the engineers are expected to move from and to both corners, creating the observed directional dispersion.

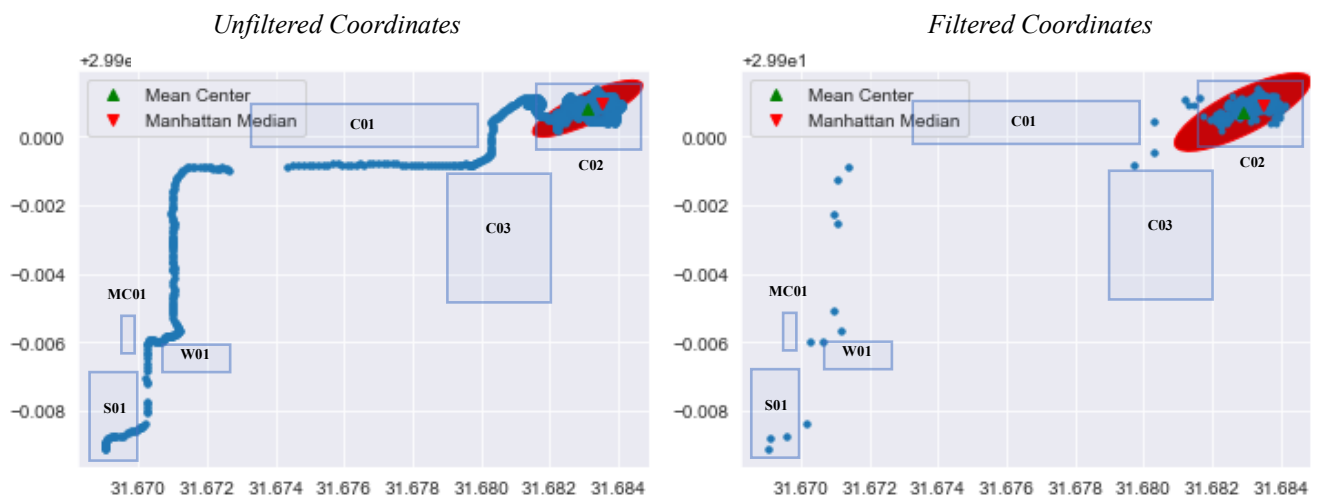


Figure 5-12: Standard Deviational Ellipse of Engineers' Coordinates

5.2.3.1.3 Spatial Randomness and Clustering

Knowing the engineers' central tendencies, their spatial randomness and clustering was tested. Firstly, applying the NND analysis yielded a p-value of $1.96588 e^{-6}$ and $1.40432 e^{-5}$ for unfiltered and filtered data, respectively. Both p-values are much lower than 0.05 signifying that the engineers are clustered and do not follow the distribution of complete spatial randomness.

To further validate the clustering of the engineers, the g-distance analysis was used. The G-function plots of the engineers are shown in Figure 5-13, and the G-function envelopes are shown in Figure 5-14.

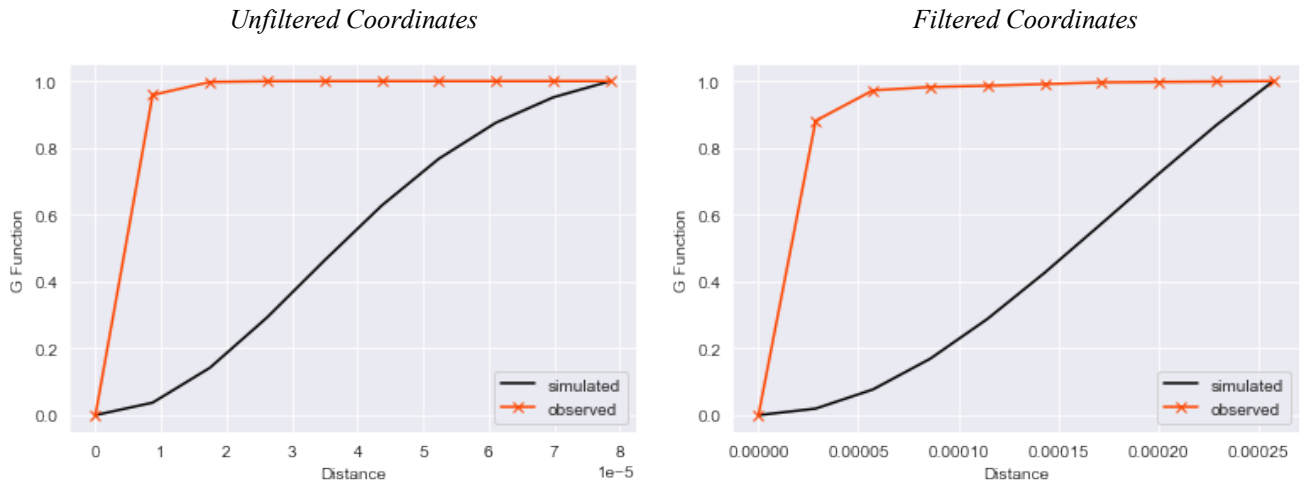


Figure 5-13: G-function Plots for Engineers' Coordinates

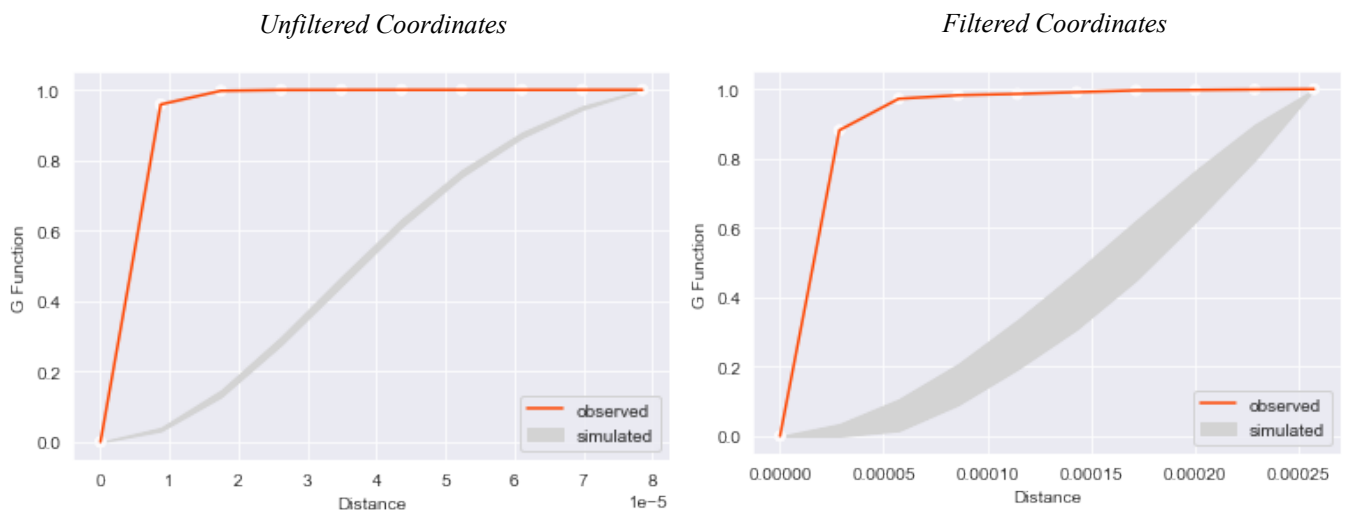


Figure 5-14: G-function Envelope Plots for Engineers' Coordinates

The G-plots confirm the spatial clustering of the engineers. As seen, the G-plots of the observed engineer coordinates is to the left of the expected G-plot of workers' following a random distribution on the construction site. The g-distance was seen to increase rapidly at shorter distances, signifying the clustering of the engineers in the area.

Given that the engineers are clustered, the BIRCH clustering technique is used to detect the number of clusters of workers on site. For this site, a single cluster was detected as shown in Figure 5-15, for both Unfiltered and filtered coordinates.

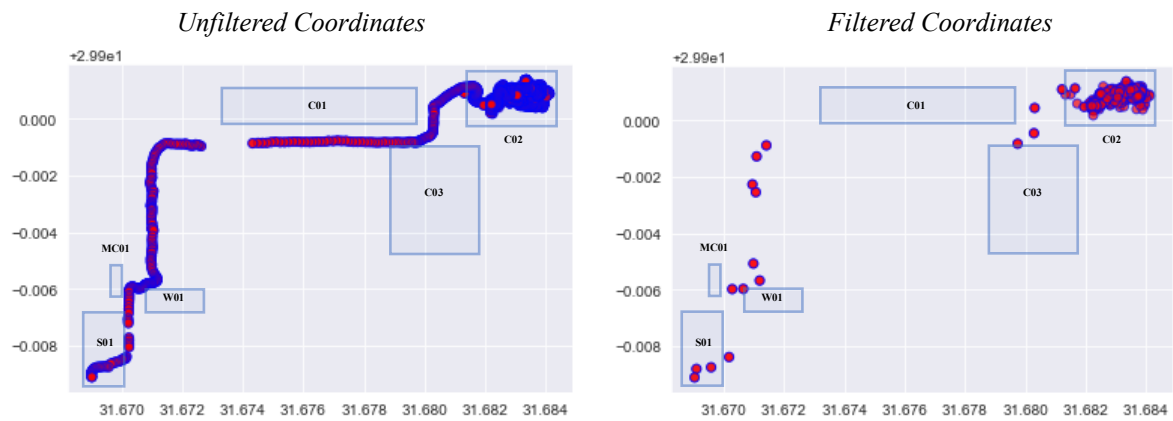


Figure 5-15: Clustering of Engineers' Coordinates

The existence of a single cluster shows that the coordinates of the engineers were close enough for them to be identified as a cluster.

Upon comparing outputs from the spatial randomness and clustering of the engineers to their observed locations, the validity of the results is proven. Given that visually it can be perceived that the engineers are heavily located in C02 over the 10-day period, then it is expected that spatial clustering is observed.

5.2.3.1.4 Workers' Density on Site

Alongside identifying the workers' clusters, workers' density was then determined using the below techniques:

a. Quadrat Analysis:

Applying quadrat analysis on the site, by dividing the site into quadrats of 10x10 yields a quadrat density count as shown in Figure 5-16. The counts are for the Unfiltered data are different from those of the filtered data, since the number coordinates is reduced, however the obtained p-value remains almost the same, $1.28439 e^{-6}$. The obtained p-value from the quadrat analysis, that is less than 0.05, confirms the spatial clustering of the engineers within the chosen number of quadrats on site, specifically in the quadrant where C02 is located, and the near-by quadrats.

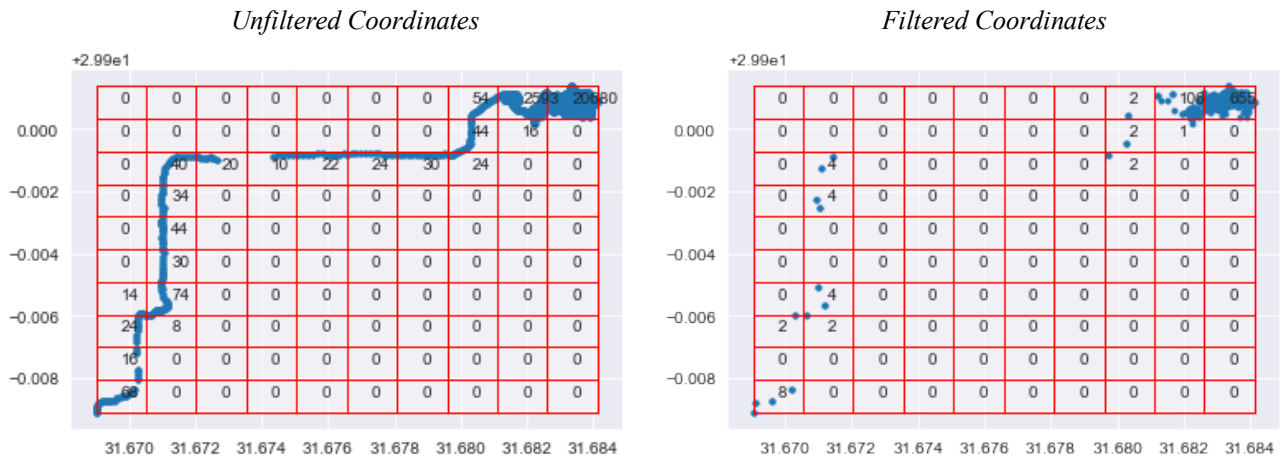


Figure 5-16: Quadrat Density of Engineers' Coordinates

b. Voronoi-based analysis:

Moreover, the Voronoi Diagram from the engineers is produced for unfiltered data, and filtered data. The Voronoi diagrams are shown in Figure 5-17.

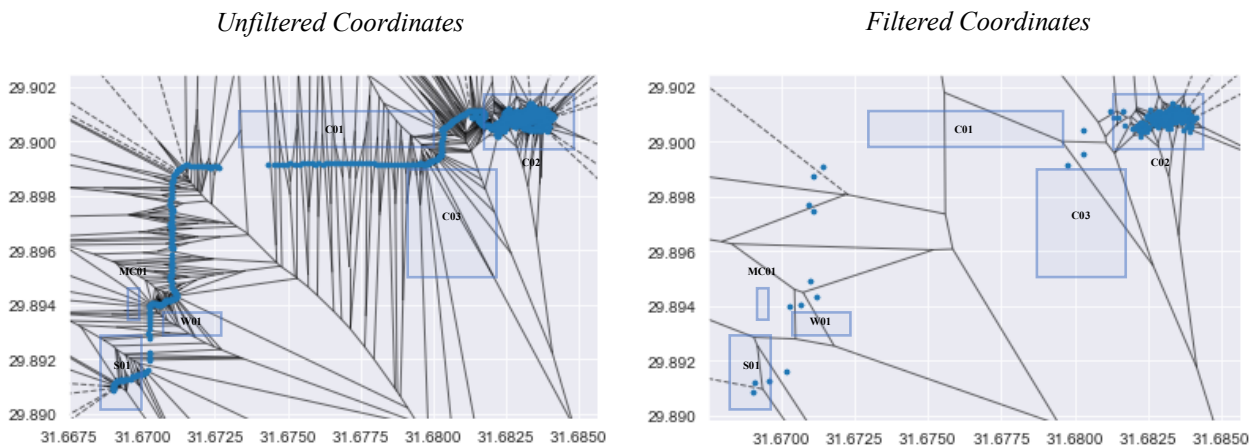


Figure 5-17: Voronoi Diagrams of Engineers' Coordinates

The visual capacity of the Voronoi cells can be observed more clearly for the filtered data. It can be depicted from the area of the Voronoi cell of the filtered data, that there was not apparent density around the engineer as he travelled from C02 to S01, this coincides with the data collected. However, the opposite was noted in the Voronoi diagram for the unfiltered data, since the Voronoi cells around the engineer in the same area indicated false densities. These false densities are due to the coordinates of the engineer being recorded every second of the time interval.

The Voronoi diagrams shown indicate the density of the engineers within the working area C02, since the Voronoi cells are particularly small.

c. Kernel Density Estimation – KDE:

Visual depiction of the engineers' densities using Voronoi analysis is then further established by using the KDE. The densities estimated were then used to plot the 2-D contoured map, also referred to as heatmap, of the engineers' densities, using a bandwidth of 0.6. The densities were used to determine the productivity zones on site. Zones of high intensity-colored contours indicate higher densities of engineers, signifying higher productivity in the zone. As the intensity becomes less, the zones are indicated to have medium to low productivity. The generated heatmaps for the Unfiltered and filtered engineers' coordinates are shown in Figure 5-18.

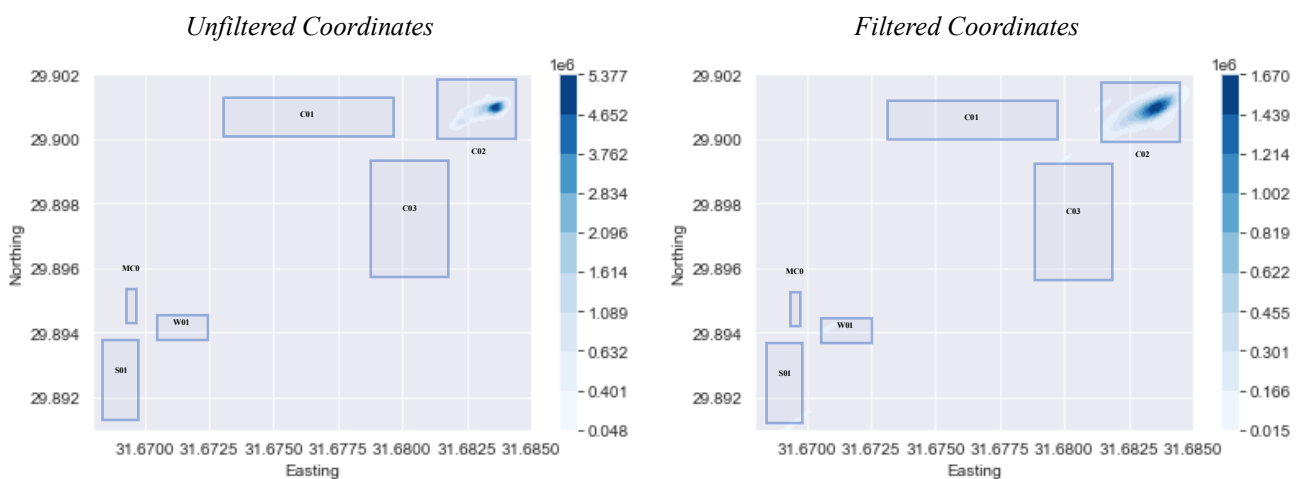


Figure 5-18: Productivity Zones of Engineers on Site

Both heatmaps have the same density scale, since the densities are relative to the total number of coordinates. Accordingly, the heatmaps point to C02 as being the highest productivity zone on site since it has the highest intensity contour of the plot. This result is as expected since the chosen engineers on site supervise area C02 only.

5.2.3.1.5 Workers' Time Distribution

Finally, the last analysis performed in the case study, is the analysis of the workers' time distribution. Given the workers location and the pre-defined zones on the site, the results of the analysis are shown in Figure 5-19.

<i>Unfiltered Coordinates</i>		<i>Filtered Coordinates</i>	
B	0.551399	B	0.571006
C02	0.447103	C02	0.424556
S01	0.001498	S01	0.004438
Name: Area Code, dtype: float64		Name: Area Code, dtype: float64	
TA	0.551399	TA	0.571006
WA	0.448601	WA	0.428994

Figure 5-19: Engineers' Time Distribution

The results for the unfiltered coordinates show that, both engineers combined, spend **44.9%** of their time in a working area, and **55.1%** traveling on site. The percentage of time spent traveling on site is higher than the time spent working for both unfiltered and filtered coordinates. However, this could be attributed to the rigidity of the calculation. There is no spatial distance around the construction area for which the worker could still be assumed to be within the working zone. The lack of tolerance contributes to the higher percentage of traveling time, given that the worker is confirmed to be traveling if located anywhere on site outside of a working or resting area.

Nevertheless, the traveling time of engineers is still relatively high than what would be favorable for a construction site. This could be attributed to the either the orientation of site layout or to the unproductivity of workers. The site chosen, however, appears to have an inefficient layout since the workshop, main caravan, and storage area are remotely located on site. This could help decision makers to make the necessary changes to improve the productivity of the site.

CHAPTER 6 – CONCLUSION & RECOMMENDATIONS

6.1 Research Summary and Overview

This research developed a framework that utilizes real-time monitoring technologies alongside spatial statistical analysis for enhanced control and monitoring of construction site performance parameters. The framework consists of three stages: (1) data collection, (2) data preparation and cleaning, and (3) data analysis. Firstly, 3 sets of data are collected, these are: (a) site geographic data – regarding geographic location of site areas, (b) site periodic data – regarding safety and quality on site, and (c) real-time geospatial data – regarding productivity. Both site geographic data and site periodic data are collected using semi-automated techniques, whereas the real-time geospatial data is collected using GPS monitoring technology. After the data is collected, the data is prepared for analysis using Python ® transformative algorithms. Finally, the prepared data is analyzed using visual and spatial temporal statistical analysis techniques using the Python ® programming language.

After the framework was developed, it was implemented on a set of data that was generated randomly. The implementation of the framework established its applicability on construction data, where the data was analyzed with the objective of investigating the performance of a construction site. The analysis was split into two phases. The first phase was the first order analysis of safety incidents, inspection requests, and workers' spatial temporal coordinates on site. The second phase was the derivate of the first order analysis, where results from first order analysis are compared and correlated against one another.

Finally, the data collection and analysis of workers' real-time geospatial data stage of the framework was applied on an actual case study for a construction project in Cairo, Egypt. Results from the case study verified the feasibility of applying the framework on construction sites. Thus, emphasizing the potential benefits gained from using real-time monitoring technologies accompanied by spatial statistical analysis of data collected from sites.

6.2 Research Contributions

The research contribution to the construction industry can be summarized as follows:

- Utilization of GPS monitoring technologies on construction sites.
- Application of spatial temporal statistical analysis of data from construction sites by using a processing algorithm that was developed using Python ®.
- Development of a framework that combines both GPS monitoring technologies and spatial temporal analysis for monitoring and control of site performance parameters.
- Verification of the applicability of the framework by implementing it on a real construction site.
- Usage of Point Pattern Analysis to investigate the conduct of construction workers on site.
- Generation of heatmaps and voronoi diagrams to visualize spatial temporal data collected from sites for prompt interpretation of data.
- Providing a novel approach that could potentially be used to unveil if and how the workers' spatial behavior affects the sites' safety and quality performance.
- Offering a distinctive method of identifying the efficiency of the site layout by analyzing the workers' spatial temporal data.
- Using the framework to establish a mechanism by which the progress on site could be monitored and compared against project baseline schedule.
- Delivering a process to assess the value of money by comparing the workers' spatial behavior with the expenditures on site.

6.3 Recommendations for Future Research

For future research, it is firstly recommended to fully automate the entire process of data collection, where a real-time approach could be utilized to collect data regarding safety incidents and inspection requests. A mobile phone application could be developed to allow users on site to input the aforementioned data in a real-time manner to enhance the data collection stage of the framework. Given that the framework was partially implemented, the second order analysis of the framework could not be utilized on an actual construction site. Accordingly, by applying the full framework on multiple construction sites, the relationships between the site parameters could be further investigated. Thus, an empirical measure of the different site parameters could be

established for improved site monitoring and control systems. Also, the elevation of the workers could be considered when studying the workers' spatial behavior as it might be a factor that could potentially have an impact on the safety, quality, and progress on site. Moreover, the physical parameters of the workers, such as oxygen levels, heartrate, calories burnt, etc. could be studied using technologies such as smart watches. The physical parameters as well as the 3-D spatial data of workers might be used to indicate the level of effort exerted by each of the workers in a specific area on site.

REFERENCES

- Becker, T. C., Jaselskis, E. J., & El-Gafy, M. (2014). Improving predictability of construction project outcomes through intentional management of indirect construction costs. *Journal of Construction Engineering and Management*.
- Calvetti, D., Meda, P., Goncalves, M. C., & Sousa, H. (2020). Worker 4.0: The future of sensed construction sites. *Buildings* (Basel), 10(10), 1.
- Chen, Yen-Chi. (2017). A Tutorial on Kernel Density Estimation and Recent Advances. *Biostatistics & Epidemiology*. 1. 10.1080/24709360.2017.1396742.
- Ferrero, M. (2011). Voronoi Diagram: The Generator Recognition Problem. *Computing Research Repository - CORR*.
- Jiang, H., Lin, P., Qiang, M., & Fan, Q. (2015). A labor consumption measurement system based on real-time tracking technology for dam construction site. *Automation in Construction*, 52, 1-15.
- Josephson, P. E., & Saukkoriipi, L. (2005). Waste in Construction Projects–Need of a Changed View. *Fou-väst*, report, 507.
- Ko, C., & Kuo, J. (2015). Making formwork construction lean. *Journal of Civil Engineering and Management*, 21(4), 444-458.
- Navon, R. (2005). Automated project performance control of construction projects. *Automation in Construction*, 14(4), 467-476.
- Navon, R., & Goldschmidt, E. (2003). Can labor inputs be measured and controlled automatically? *Journal of Construction Engineering and Management*, 129(4), 437-445.
- Nikakhtar, A., Hosseini, A. A., Wong, K. Y., & Zavichi, A. (2015). Application of lean construction principles to reduce construction process waste using computer simulation: a case study. *International Journal of Services and Operations Management*, 20(4), 461480.
- Ramadhani, Fanny & Zarlis, Muhammad & Suwilo, Saib. (2020). Improve BIRCH algorithm for big data clustering. *IOP Conference Series: Materials Science and Engineering*. 725. 012090. 10.1088/1757-899X/725/1/012090.
- Soltanmohammadlou, N., Sadeghi, S., Hon, C. K. H., & Mokhtarpour-Khanghah, F. (2019). Real-time locating systems and safety in construction sites: A literature review. *Safety Science*, 117, 229-242.

- Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, 29(2), 225-238.
- Thomas, H. R., Horman, M. J., Minchin, R. E., & Chen, D. (2003). Improving labor flow reliability for better productivity as lean construction principle. *Journal of Construction Engineering and Management*, 129(3), 251-261.
- Wang, T., Yang, T., Yang, C., & Chan, F. T. S. (2015). Lean principles and simulation optimization for emergency department layout design. *Industrial Management & Data Systems*, 115(4), 678-699.
- Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open-Source Software*, 6(60), 3021, <https://doi.org/10.21105/joss.03021>.
- Yuan, Y., Qiang, Y., Bin Asad, K., and Chow, T. E. (2020). Point Pattern Analysis. *The Geographic Information Science & Technology Body of Knowledge* (1st Quarter 2020 Edition), John P. Wilson (ed.).
- Zhang, S., Shang, C., Fang, X., He, S., Yu, L., Wang, C., & Yan, L. (2021). Wireless Monitoring–Based real-time analysis and early-warning safety system for deep and large underground caverns. *Journal of Performance of Constructed Facilities*, 35(2).

APPENDIX A – PYTHON ALGORITHM

```
In [1]: import pandas as pd
import csv
import os
import geopandas as gpd
import folium
from shapely.geometry import Point, Polygon, LinearRing, LineString
import shapely
from shapely.ops import cascaded_union, unary_union, split
import numpy as np
import matplotlib.pyplot as plt
import pycrs
import fiona
import seaborn as sns
from pyproj import CRS
from fiona.crs import from_epsg
from sklearn.cluster import KMeans

sd = pd.read_csv('/Users/hoda/Desktop/Thesis/Safety_Trial.csv')

rd = pd.read_csv('/Users/hoda/Desktop/Thesis/IR_Trial.csv')

bd1 = pd.read_csv('/Users/hoda/Desktop/Thesis/Trial_2.csv')

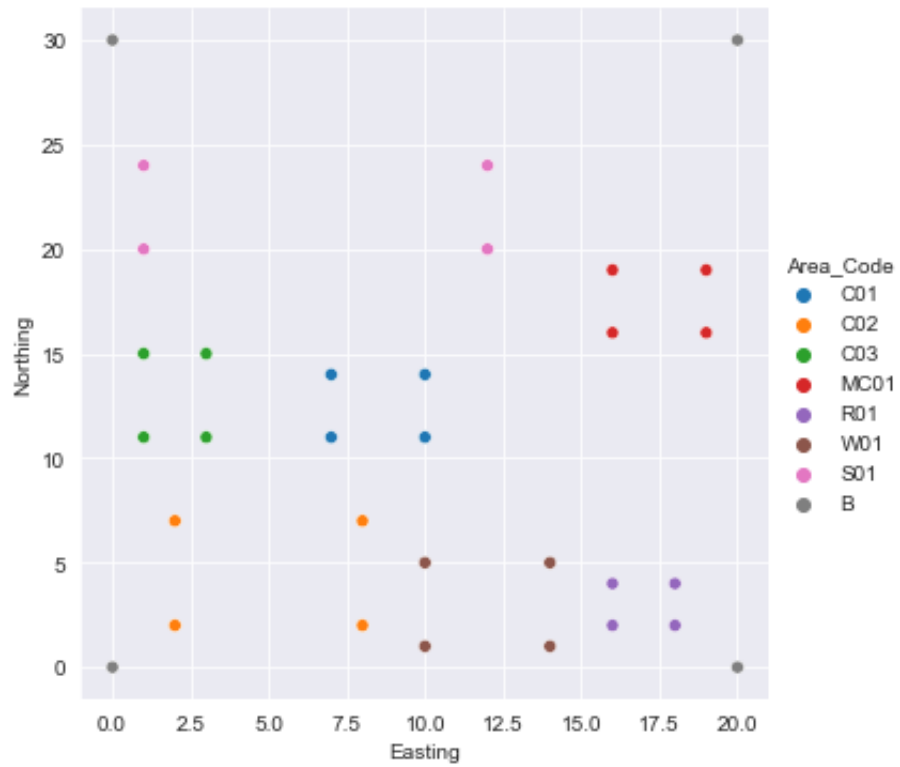
sd
```

```
Out[1]:
```

	Area_Code	No_Incidents	Date	Time
0	C01	5.0	12/15/20	14:17
1	C02	6.0	12/16/20	09:58
2	C03	3.0	12/17/20	09:16
3	C01	8.0	12/18/20	09:41
4	C02	7.0	12/19/20	09:33
5	C03	5.0	12/20/20	15:25
6	C01	3.0	12/21/20	15:26
7	C02	8.0	12/22/20	08:56
8	C03	4.0	12/23/20	10:12
9	C01	5.0	12/24/20	14:59
10	C02	3.0	12/25/20	11:54
11	C03	2.0	12/26/20	12:02
12	C01	5.0	12/27/20	09:52
13	C02	4.0	12/28/20	09:56
14	C01	7.0	12/29/20	11:06

```
In [2]: sns.set_style('darkgrid')
sns.relplot(data=bd1, x="Easting", y="Northing", hue="Area_Code")
```


Out[2]: <seaborn.axisgrid.FacetGrid at 0x10a521190>



```

In [3]: bd = gpd.GeoDataFrame(bd1, crs="EPSG:4326", geometry = gpd.points_from_xy(bd1.x, bd1.y))

bd = bd.drop('Ref._No.', 1)
UniqueNames = bd.Area_Code.unique()
bd_of_Codes = pd.DataFrame(UniqueNames)

DataFrameDict = {elem : pd.DataFrame for elem in UniqueNames}

for key in DataFrameDict.keys():
    DataFrameDict[key] = bd[:, bd.Area_Code == key]

bd_of_Codes.columns = ['Area_Code']

for count, blg in enumerate(UniqueNames):
    i = str(blg)
    bd = DataFrameDict[blg].iloc[:, 1:3]
    Coordinates = bd.to_records(index=False)
    list_coordinates = list(Coordinates)
    globals()[blg]=Polygon(list_coordinates)

shape = []
for count, blg in enumerate(UniqueNames):
    shape.append(eval(blg))

bd_gdf = pd.DataFrame(shape)
bd_gdf.columns = ['Polygon']
bd_shapes = pd.concat([bd_of_Codes, bd_gdf], axis=1)
bd_shapes = gpd.GeoDataFrame(bd_shapes, crs="EPSG:4326", geometry='Polygon')

bd = pd.merge(bd_shapes, bd1, 'inner', 'Area_Code')
bd = bd.drop_duplicates('Polygon')
bd = bd.drop('Easting', 1)
bd = bd.drop('Northing', 1)
bd = bd.drop('Elevation', 1)
bd = bd.drop('geometry', 1)

bd

```

```

Out[3]:

```

	Area_Code	Polygon	Ref._No.	Category
0	C01	POLYGON ((10.00000 11.00000, 7.00000 11.00000, ...	1	WA
4	C02	POLYGON ((2.00000 2.00000, 2.00000 7.00000, 8....	5	WA
8	C03	POLYGON ((3.00000 11.00000, 1.00000 11.00000, ...	9	WA
12	MC01	POLYGON ((16.00000 16.00000, 16.00000 19.00000...	13	WA
16	R01	POLYGON ((16.00000 2.00000, 16.00000 4.00000, ...	17	RA
20	W01	POLYGON ((10.00000 1.00000, 10.00000 5.00000, ...	21	WA
24	S01	POLYGON ((1.00000 20.00000, 1.00000 24.00000, ...	25	WA
28	B	POLYGON ((0.00000 0.00000, 0.00000 30.00000, 2...	29	TA

```
In [4]: bd_c = bd.copy()

bd_c.geometry = bd_c['Polygon'].centroid

bd_c.crs = bd.crs
bd_c.head()

bd_c['Easting'] = bd_c['Polygon'].x
bd_c['Northing'] = bd_c['Polygon'].y

bd_c = bd_c.drop('Polygon', 1)
bd_c = bd_c.drop('Category', 1)

bd_c

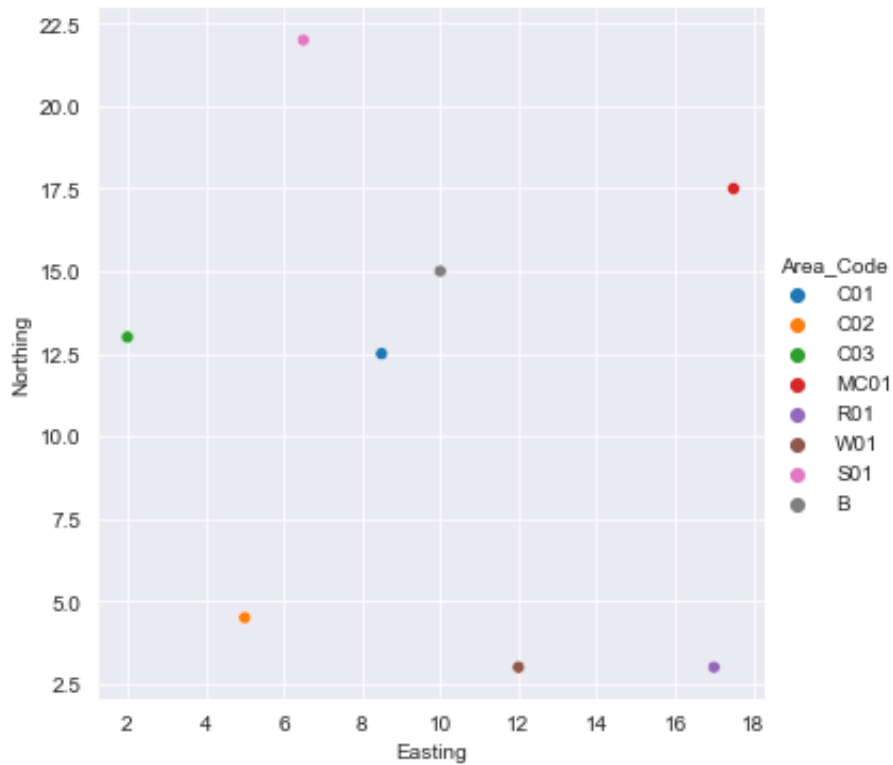
sns.relplot(data=bd_c, x="Easting", y="Northing", hue="Area_Code")
bd_c
```

```
<ipython-input-4-50623e23ed17>:3: UserWarning: Geometry is in a geographic CRS. Results from 'centroid' are likely incorrect. Use 'GeoSeries.to_crs()' to re-project geometries to a projected CRS before this operation.
```

```
bd_c.geometry = bd_c['Polygon'].centroid
```

Out[4]:

	Area_Code	Ref.No.	Easting	Northing
0	C01	1	8.5	12.5
4	C02	5	5.0	4.5
8	C03	9	2.0	13.0
12	MC01	13	17.5	17.5
16	R01	17	17.0	3.0
20	W01	21	12.0	3.0
24	S01	25	6.5	22.0
28	B	29	10.0	15.0



```

In [5]: import numpy as np
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import matplotlib.cbook as cbook

sd['avg'] = sd.No_Incidents.rolling(3).mean()

sd[['Date', 'No_Incidents', 'avg']].head()

sns.set_style('darkgrid')

fig, ax = plt.subplots()
ax.plot('Date', 'No_Incidents', data=sd, label = "Daily")
ax.plot('Date', 'avg', data=sd, color = "Red", label = "3 Day Moving Avg.")

fmt_half_year = mdates.DayLocator(interval=3)
ax.xaxis.set_major_locator(fmt_half_year)

# Minor ticks every month.
fmt_month = mdates.DayLocator()
ax.xaxis.set_minor_locator(fmt_month)

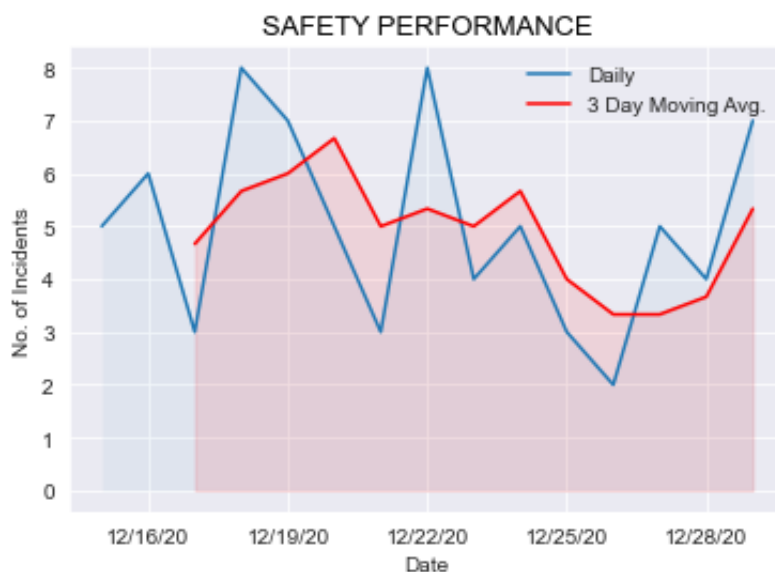
plt.fill_between(sd.Date.values, sd.avg.values, color = "Red", alpha=0.1)
plt.fill_between(sd.Date.values, sd.No_Incidents.values, alpha=0.05)

plt.xlabel("Date")
plt.ylabel("No. of Incidents")
plt.title(label="SAFETY PERFORMANCE", fontsize=13,
          color="Black")

ax.legend(loc='upper right', frameon=False)

```

Out[5]: <matplotlib.legend.Legend at 0x123019b80>



```
In [6]: sd_agg = sd.agg( ['mean' , 'max' , 'min' , 'median' , 'std'] )
sd_agg = sd_agg.replace(np.nan, '-', regex=True)

sd_agg
```

```
Out[6]:
```

	Area_Code	No_Incidents	Date	Time	avg
max	C03	8.000000	12/29/20	15:26	6.666667
min	C01	2.000000	12/15/20	08:56	3.333333
mean	-	5.000000	-	-	4.897436
median	-	5.000000	-	-	5.000000
std	-	1.889822	-	-	1.048673

```
In [16]: sd_c = pd.DataFrame.merge(sd,bd_c,on='Area_Code')

sd_c = sd_c.drop('avg', 1)
sd_c = sd_c.drop('Ref._No.', 1)
sd_c = sd_c.drop('Time', 1)

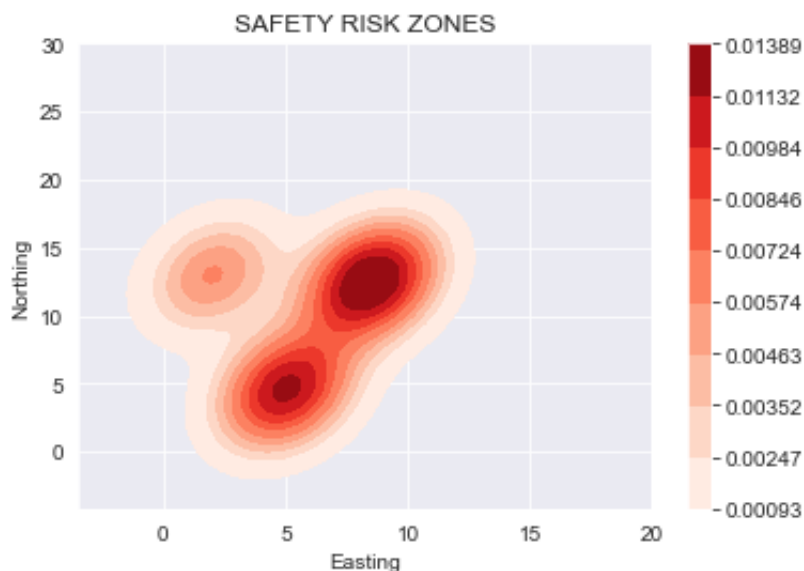
sd_kde = sd_c.reindex(sd_c.index.repeat(sd_c.No_Incidents))

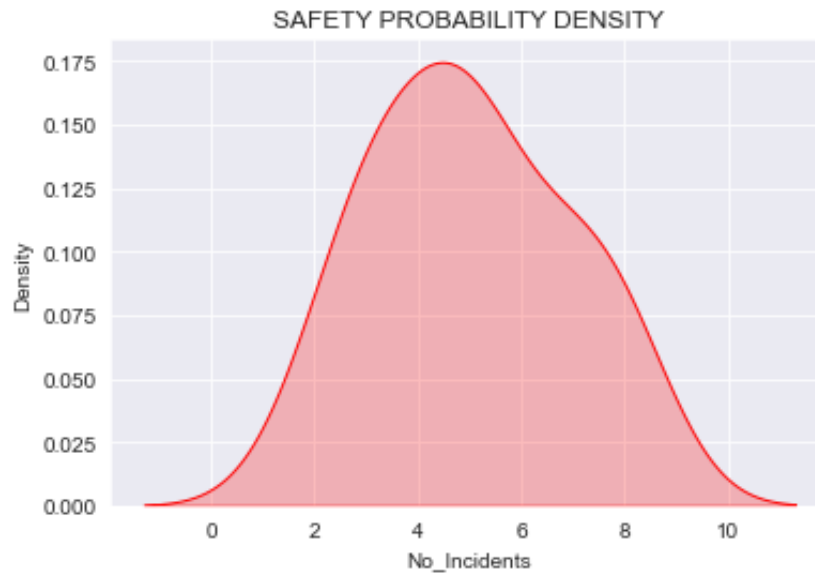
sns.kdeplot(x=sd_kde.Easting, y=sd_kde.Northing, cbar = True, cmap="Reds",
plt.xticks([0, 5, 10, 15, 20])
plt.yticks([0, 5, 10, 15, 20,25,30])

plt.show()

sns.kdeplot(sd_c.No_Incidents,color='Red',shade="Red").set(title='SAFETY PI
plt.show()

sd_c
```





Out[16]:

	Area_Code	No_Incidents	Date	Easting	Northing
0	C01	5.0	12/15/20	8.5	12.5
1	C01	8.0	12/18/20	8.5	12.5
2	C01	3.0	12/21/20	8.5	12.5
3	C01	5.0	12/24/20	8.5	12.5
4	C01	5.0	12/27/20	8.5	12.5
5	C01	7.0	12/29/20	8.5	12.5
6	C02	6.0	12/16/20	5.0	4.5
7	C02	7.0	12/19/20	5.0	4.5
8	C02	8.0	12/22/20	5.0	4.5
9	C02	3.0	12/25/20	5.0	4.5
10	C02	4.0	12/28/20	5.0	4.5
11	C03	3.0	12/17/20	2.0	13.0
12	C03	5.0	12/20/20	2.0	13.0
13	C03	4.0	12/23/20	2.0	13.0
14	C03	2.0	12/26/20	2.0	13.0

```
In [8]: rd['Acc_avg'] = rd.Accepted.rolling(3).mean()

rd[['Date', 'Accepted', 'Acc_avg']].head()

sns.set_style('darkgrid')

fig, ax = plt.subplots()
ax.plot('Date', 'Accepted', data=rd, label = "Accepted - Daily")
ax.plot('Date', 'Acc_avg', data=rd, color = "Green", label = "Accepted - 3 I

fmt_half_year = mdates.DayLocator(interval=3)
ax.xaxis.set_major_locator(fmt_half_year)

fmt_month = mdates.DayLocator()
ax.xaxis.set_minor_locator(fmt_month)

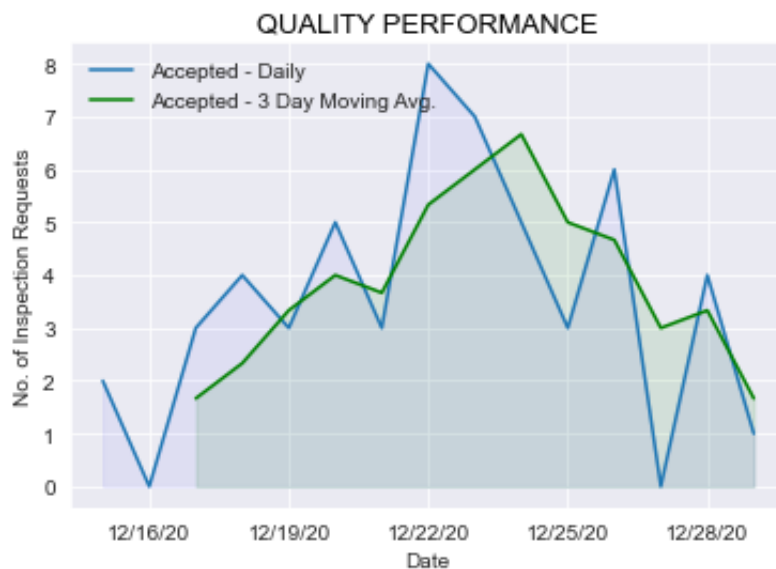
plt.fill_between(rd.Date.values, rd.Acc_avg.values, color = "Green", alpha=0.5)
plt.fill_between(rd.Date.values, rd.Accepted.values, color = "Blue", alpha=0.5)

plt.xlabel("Date")
plt.ylabel("No. of Inspection Requests")
plt.title(label="QUALITY PERFORMANCE", fontsize=13,
          color="Black")

ax.legend(loc='upper left', frameon=False)
rd
```


Out[8]:

	Area_Code	Rejected	Accepted	Date	Acc_avg
0	C01	10.0	2.0	12/15/20	NaN
1	C02	9.0	0.0	12/16/20	NaN
2	C03	2.0	3.0	12/17/20	1.666667
3	C01	1.0	4.0	12/18/20	2.333333
4	C02	3.0	3.0	12/19/20	3.333333
5	C03	9.0	5.0	12/20/20	4.000000
6	C01	10.0	3.0	12/21/20	3.666667
7	C02	12.0	8.0	12/22/20	5.333333
8	C03	4.0	7.0	12/23/20	6.000000
9	C01	4.0	5.0	12/24/20	6.666667
10	C02	4.0	3.0	12/25/20	5.000000
11	C03	5.0	6.0	12/26/20	4.666667
12	C01	7.0	0.0	12/27/20	3.000000
13	C02	8.0	4.0	12/28/20	3.333333
14	C01	10.0	1.0	12/29/20	1.666667



```
In [9]: rd['Rej_avg'] = rd.Accepted.rolling(3).mean()

rd[['Date', 'Rejected', 'Rej_avg']].head()

fig, ax = plt.subplots()

ax.plot('Date', 'Rejected', data=rd, color = "Purple", label = "Rejected - D")
ax.plot('Date', 'Rej_avg', data=rd, color = "Orange", label = "Rejected - 3")

fmt_half_year = mdates.DayLocator(interval=3)
ax.xaxis.set_major_locator(fmt_half_year)

fmt_month = mdates.DayLocator()
ax.xaxis.set_minor_locator(fmt_month)

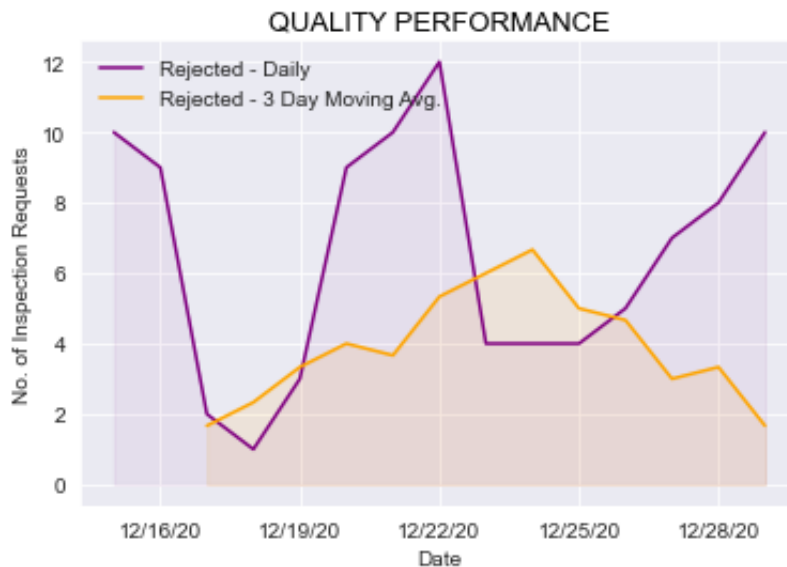
plt.fill_between(rd.Date.values, rd.Rej_avg.values, color = "Orange", alpha=0.5)
plt.fill_between(rd.Date.values, rd.Rejected.values, color = "Purple", alpha=0.5)

plt.xlabel("Date")
plt.ylabel("No. of Inspection Requests")
plt.title(label="QUALITY PERFORMANCE", fontsize=13,
          color="Black")

ax.legend(loc='upper left', frameon=False)
rd
```

Out[9]:

	Area_Code	Rejected	Accepted	Date	Acc_avg	Rej_avg
0	C01	10.0	2.0	12/15/20	NaN	NaN
1	C02	9.0	0.0	12/16/20	NaN	NaN
2	C03	2.0	3.0	12/17/20	1.666667	1.666667
3	C01	1.0	4.0	12/18/20	2.333333	2.333333
4	C02	3.0	3.0	12/19/20	3.333333	3.333333
5	C03	9.0	5.0	12/20/20	4.000000	4.000000
6	C01	10.0	3.0	12/21/20	3.666667	3.666667
7	C02	12.0	8.0	12/22/20	5.333333	5.333333
8	C03	4.0	7.0	12/23/20	6.000000	6.000000
9	C01	4.0	5.0	12/24/20	6.666667	6.666667
10	C02	4.0	3.0	12/25/20	5.000000	5.000000
11	C03	5.0	6.0	12/26/20	4.666667	4.666667
12	C01	7.0	0.0	12/27/20	3.000000	3.000000
13	C02	8.0	4.0	12/28/20	3.333333	3.333333
14	C01	10.0	1.0	12/29/20	1.666667	1.666667



```
In [10]: rd_agg = rd.agg( ['mean' , 'max' , 'min', 'median', 'std'] )
rd_agg = rd_agg.replace(np.nan, '-', regex=True)

rd_agg
```

Out[10]:

	Area_Code	Rejected	Accepted	Date	Acc_avg	Rej_avg
max	C03	12.000000	8.000000	12/29/20	6.666667	6.666667
min	C01	1.000000	0.000000	12/15/20	1.666667	1.666667
mean	-	6.533333	3.600000	-	3.897436	3.897436
median	-	7.000000	3.000000	-	3.666667	3.666667
std	-	3.440653	2.354327	-	1.577531	1.577531

In [20]:

```

rd_c = pd.DataFrame.merge(rd,bd_c,on='Area_Code')

rd_c = rd_c.drop('Acc_avg', 1)
rd_c = rd_c.drop('Ref_No.', 1)
rd_c = rd_c.drop('Rej_avg', 1)
rd_c_Acc = rd_c.drop('Accepted', 1)
rd_c_Rej = rd_c.drop('Rejected', 1)

rd_c_Acc_kde = rd_c_Acc.reindex(rd_c_Acc.index.repeat(rd_c_Acc.Rejected))
rd_c_Rej_kde = rd_c_Rej.reindex(rd_c_Rej.index.repeat(rd_c_Rej.Accepted))

sns.kdeplot(x=rd_c_Acc_kde.Easting, y=rd_c_Acc_kde.Northing, cbar = True,cr
plt.xticks([0, 5, 10, 15, 20])
plt.yticks([0, 5, 10, 15, 20,25,30])
plt.show()

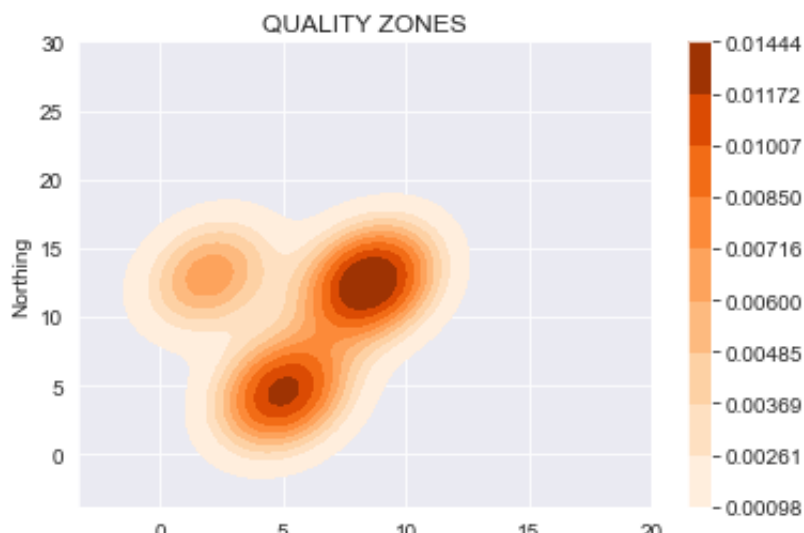
sns.kdeplot(x=rd_c_Rej_kde .Easting, y=rd_c_Rej_kde .Northing, cbar = True
plt.xticks([0, 5, 10, 15, 20])
plt.yticks([0, 5, 10, 15, 20,25,30])
plt.show()

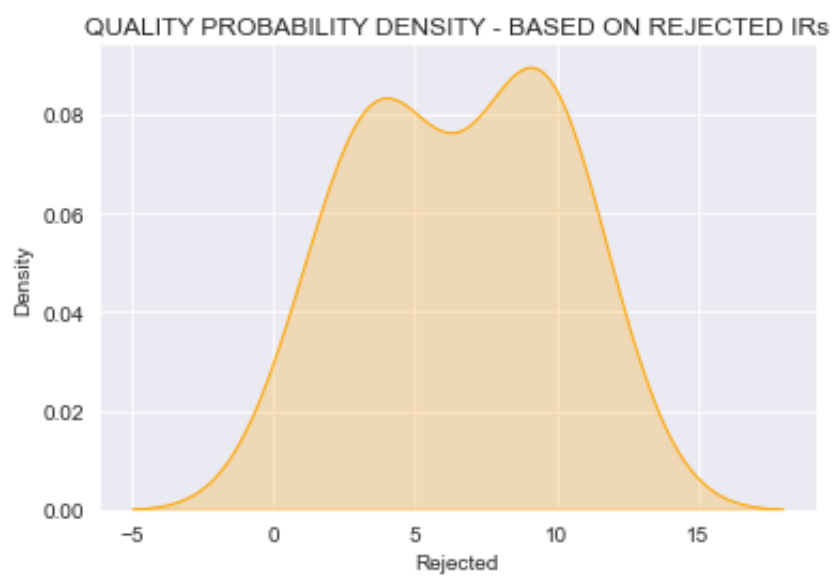
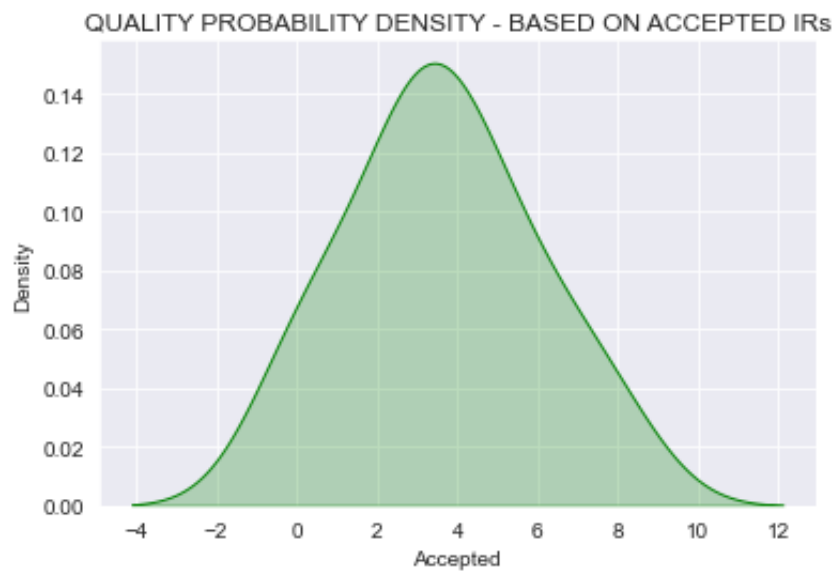
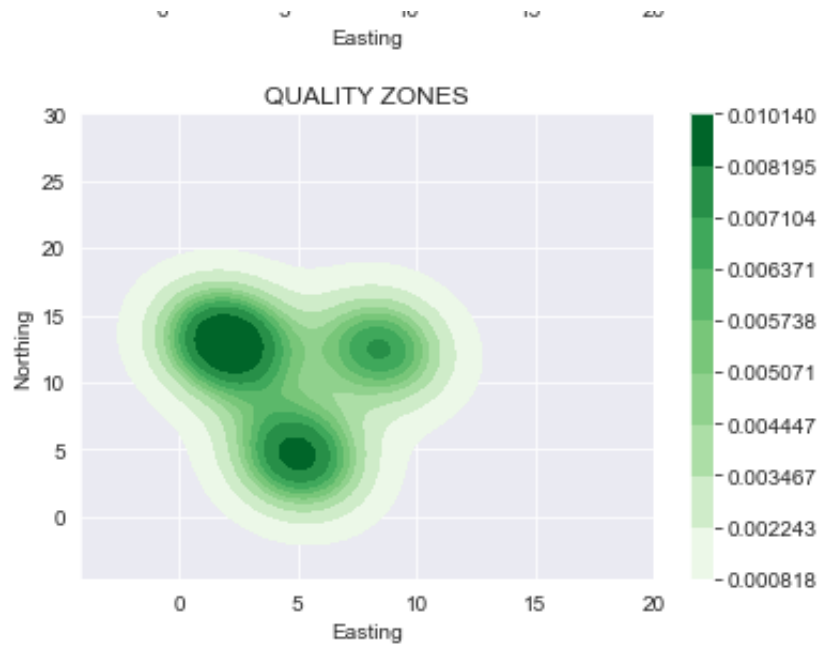
sns.kdeplot(rd_c.Accepted,color='Green',shade="Green").set(title='QUALITY I
plt.show()

sns.kdeplot(rd_c.Rejected,color='Orange',shade="Orange").set(title='QUALITY
plt.show()

rd_c_Rej

```





Out[20]:

	Area_Code	Accepted	Date	Easting	Northing
0	C01	2.0	12/15/20	8.5	12.5
1	C01	4.0	12/18/20	8.5	12.5
2	C01	3.0	12/21/20	8.5	12.5
3	C01	5.0	12/24/20	8.5	12.5
4	C01	0.0	12/27/20	8.5	12.5
5	C01	1.0	12/29/20	8.5	12.5
6	C02	0.0	12/16/20	5.0	4.5
7	C02	3.0	12/19/20	5.0	4.5
8	C02	8.0	12/22/20	5.0	4.5
9	C02	3.0	12/25/20	5.0	4.5
10	C02	4.0	12/28/20	5.0	4.5
11	C03	3.0	12/17/20	2.0	13.0
12	C03	5.0	12/20/20	2.0	13.0
13	C03	7.0	12/23/20	2.0	13.0
14	C03	6.0	12/26/20	2.0	13.0

```
In [1]: import pandas as pd
import csv
import os
import geopandas as gpd
import folium
from shapely.geometry import Point, Polygon, LinearRing, LineString
import shapely
from shapely.ops import cascaded_union, unary_union, split
import numpy as np
import matplotlib.pyplot as plt
import pyproj
import fiona
import seaborn as sns
from pyproj import CRS
from fiona.crs import from_epsg
from sklearn.cluster import KMeans

wd = pd.read_csv('/Users/hoda/Desktop/Thesis/Trial_Data.csv')

bd1 = pd.read_csv('/Users/hoda/Desktop/Thesis/Trial_2.csv')

bd1
```

Out[1]:

	Area_Code	Ref._No.	Easting	Northing	Elevation	Category
0	C01	1	10.0	11.0	412.0	WA
1	C01	2	7.0	11.0	412.0	WA
2	C01	3	7.0	14.0	412.0	WA
3	C01	4	10.0	14.0	412.0	WA
4	C02	5	2.0	2.0	412.0	WA
5	C02	6	2.0	7.0	412.0	WA
6	C02	7	8.0	7.0	412.0	WA
7	C02	8	8.0	2.0	412.0	WA
8	C03	9	3.0	11.0	412.0	WA
9	C03	10	1.0	11.0	412.0	WA
10	C03	11	1.0	15.0	412.0	WA
11	C03	12	3.0	15.0	412.0	WA
12	MC01	13	16.0	16.0	412.0	WA
13	MC01	14	16.0	19.0	412.0	WA
14	MC01	15	19.0	19.0	412.0	WA
15	MC01	16	19.0	16.0	412.0	WA
16	R01	17	16.0	2.0	412.0	RA
17	R01	18	16.0	4.0	412.0	RA
18	R01	19	18.0	4.0	412.0	RA
19	R01	20	18.0	2.0	412.0	RA
20	W01	21	10.0	1.0	412.0	WA
21	W01	22	10.0	5.0	412.0	WA
22	W01	23	14.0	5.0	412.0	WA
23	W01	24	14.0	1.0	412.0	WA
24	S01	25	1.0	20.0	412.0	WA
25	S01	26	1.0	24.0	412.0	WA
26	S01	27	12.0	24.0	412.0	WA
27	S01	28	12.0	20.0	412.0	WA
28	B	29	0.0	0.0	412.0	TA
29	B	30	0.0	30.0	412.0	TA
30	B	31	20.0	30.0	412.0	TA
31	B	32	20.0	0.0	412.0	TA


```

In [2]: bd = gpd.GeoDataFrame(bd1, crs="EPSG:4326", geometry = gpd.points_from_xy(bd1.x, bd1.y))

bd = bd.drop('Ref._No.', 1)
UniqueNames = bd.Area_Code.unique()
bd_of_Codes = pd.DataFrame(UniqueNames)

DataFrameDict = {elem : pd.DataFrame for elem in UniqueNames}

for key in DataFrameDict.keys():
    DataFrameDict[key] = bd[:, [bd.Area_Code == key]]

bd_of_Codes.columns = ['Area_Code']

for count, blg in enumerate(UniqueNames):
    i = str(blg)
    bd = DataFrameDict[blg].iloc[:, 1:3]
    Coordinates = bd.to_records(index=False)
    list_coordinates = list(Coordinates)
    globals()[blg]=Polygon(list_coordinates)

shape = []
for count, blg in enumerate(UniqueNames):
    shape.append(eval(blg))

bd_gdf = pd.DataFrame(shape)
bd_gdf.columns = ['Polygon']
bd_shapes = pd.concat([bd_of_Codes, bd_gdf], axis=1)
bd_shapes = gpd.GeoDataFrame(bd_shapes, crs="EPSG:4326", geometry='Polygon')

bd = pd.merge(bd_shapes, bd1, 'inner', 'Area_Code')
bd = bd.drop_duplicates('Polygon')
bd = bd.drop('Easting', 1)
bd = bd.drop('Northing', 1)
bd = bd.drop('Elevation', 1)
bd = bd.drop('geometry', 1)

bd

bd2 = bd[bd.Area_Code != 'B']

bd2

```

Out[2]:

	Area_Code	Polygon	Ref_No.	Category
0	C01	POLYGON ((10.00000 11.00000, 7.00000 11.00000,...	1	WA
4	C02	POLYGON ((2.00000 2.00000, 2.00000 7.00000, 8....	5	WA
8	C03	POLYGON ((3.00000 11.00000, 1.00000 11.00000, ...	9	WA
12	MC01	POLYGON ((16.00000 16.00000, 16.00000 19.00000...	13	WA
16	R01	POLYGON ((16.00000 2.00000, 16.00000 4.00000, ...	17	RA
20	W01	POLYGON ((10.00000 1.00000, 10.00000 5.00000, ...	21	WA
24	S01	POLYGON ((1.00000 20.00000, 1.00000 24.00000, ...	25	WA

In [3]:

```

bd_c = bd.copy()

bd_c.geometry = bd_c['Polygon'].centroid

bd_c.crs = bd.crs
bd_c.head()

bd_c['Easting'] = bd_c['Polygon'].x
bd_c['Northing'] = bd_c['Polygon'].y

bd_c = bd_c.drop('Polygon', 1)

bd_c2 = bd_c[bd_c.Area_Code != 'B']

bd_c2

```

<ipython-input-3-bdd36bce6e7f>:3: UserWarning: Geometry is in a geographic CRS. Results from 'centroid' are likely incorrect. Use 'GeoSeries.to_crs()' to re-project geometries to a projected CRS before this operation.

Out[3]:

```

bd_c.geometry = bd_c['Polygon'].centroid

```

	Area_Code	Ref_No.	Category	Easting	Northing
0	C01	1	WA	8.5	12.5
4	C02	5	WA	5.0	4.5
8	C03	9	WA	2.0	13.0
12	MC01	13	WA	17.5	17.5
16	R01	17	RA	17.0	3.0
20	W01	21	WA	12.0	3.0
24	S01	25	WA	6.5	22.0

```
In [4]: sns.set_style('darkgrid')

#sns.relplot(data=bd1,x="Easting", y="Northing", hue="Area_Code")

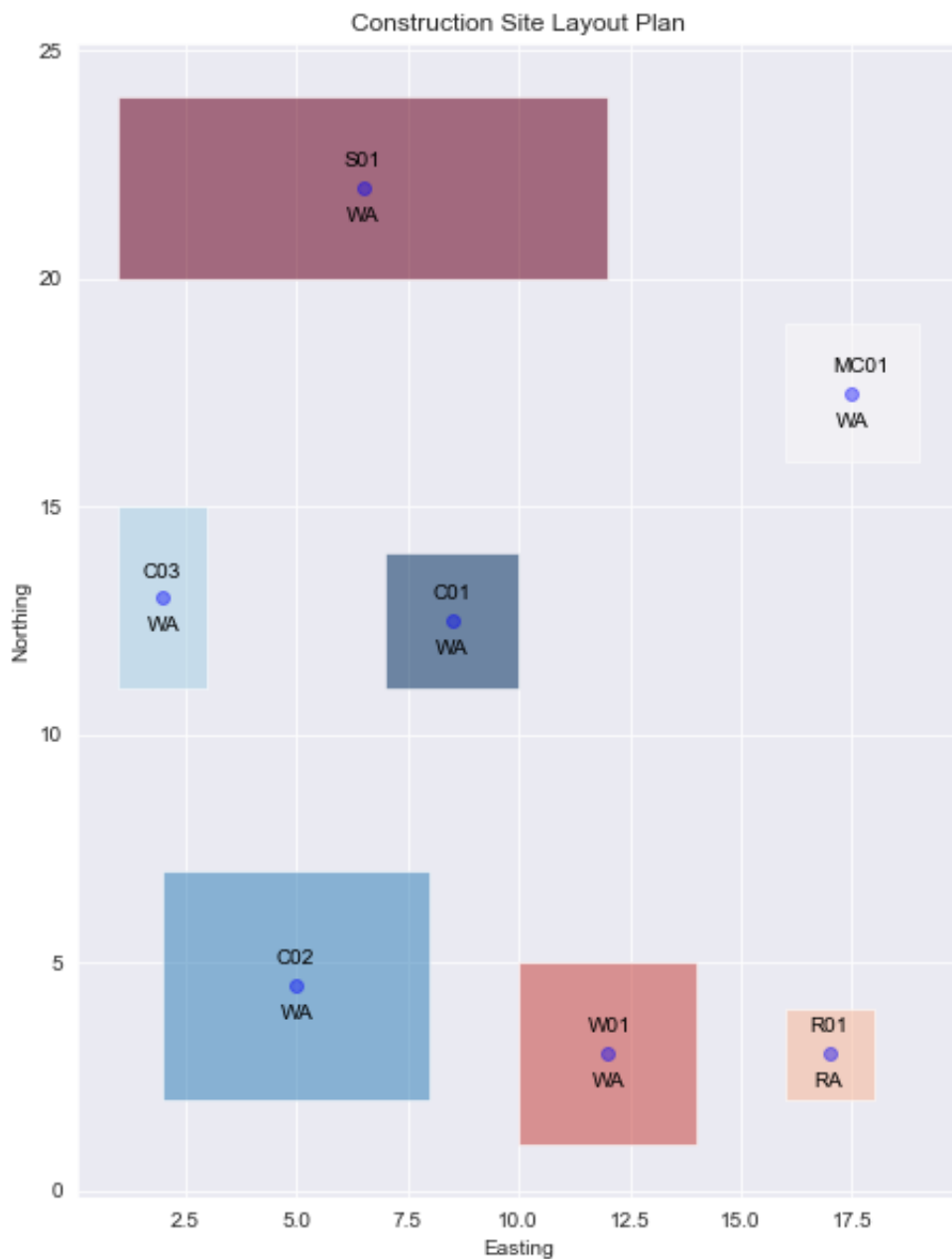
bd2.plot(figsize=(10, 10), alpha=0.55, cmap='RdBu_r')

plt.scatter(bd_c2.Easting, bd_c2.Northing, alpha=0.4, c='Blue')

for i, point in bd_c2.iterrows():

    plt.text(point['Easting']-0.45, point['Northing']+0.45, str(point['Area
    plt.text(point['Easting']-0.35, point['Northing']-0.75, str(point['Cate

plt.title("Construction Site Layout Plan")
plt.xlabel("Easting")
plt.ylabel("Northing")
plt.show()
```



```

In [5]: wd_trips = gpd.GeoDataFrame(wd, geometry = gpd.points_from_xy(wd['Easting',
wd_trips = gpd.GeoDataFrame(wd_trips, crs="EPSG:4326", geometry='geometry')

wd_joined = gpd.sjoin(wd_trips, bd, 'inner', op='within')

wd_joined = wd_joined.drop('index_right', 1)
wd_joined = wd_joined.drop('Ref._No.', 1)

bd = gpd.GeoDataFrame(bd1, crs="EPSG:4326", geometry = gpd.points_from_xy(bd1['Easting',
bd = bd.drop('Ref._No.', 1)
UniqueNames = bd.Area_Code.unique()
bd_of_Codes = pd.DataFrame(UniqueNames)

DataFrameDict = {elem : pd.DataFrame for elem in UniqueNames}

for key in DataFrameDict.keys():
    DataFrameDict[key] = bd[bd.Area_Code == key]

bd_of_Codes.columns = ['Area_Code']

for count, blg in enumerate(UniqueNames):
    i = str(blg)
    bd = DataFrameDict[blg].iloc[:, 1:3]
    Coordinates = bd.to_records(index=False)
    list_coordinates = list(Coordinates)
    globals()[blg]=Polygon(list_coordinates)

shape = []
for count, blg in enumerate(UniqueNames):
    shape.append(eval(blg))

bd_gdf = pd.DataFrame(shape)
bd_gdf.columns = ['Polygon']
bd_shapes = pd.concat([bd_of_Codes, bd_gdf], axis=1)
bd_shapes = gpd.GeoDataFrame(bd_shapes, crs="EPSG:4326", geometry='Polygon')

bd = pd.merge(bd_shapes, bd1, 'inner', 'Area_Code')
bd = bd.drop_duplicates('Polygon')
bd = bd.drop('Easting', 1)
bd = bd.drop('Northing', 1)
bd = bd.drop('Elevation', 1)
bd = bd.drop('geometry', 1)

bd

bd2 = bd[bd.Area_Code != 'B']

bd2

```

Out[5]:

	Area_Code	Polygon	Ref_No.	Category
0	C01	POLYGON ((10.00000 11.00000, 7.00000 11.00000,...	1	WA
4	C02	POLYGON ((2.00000 2.00000, 2.00000 7.00000, 8....	5	WA
8	C03	POLYGON ((3.00000 11.00000, 1.00000 11.00000, ...	9	WA
12	MC01	POLYGON ((16.00000 16.00000, 16.00000 19.00000...	13	WA
16	R01	POLYGON ((16.00000 2.00000, 16.00000 4.00000, ...	17	RA
20	W01	POLYGON ((10.00000 1.00000, 10.00000 5.00000, ...	21	WA
24	S01	POLYGON ((1.00000 20.00000, 1.00000 24.00000, ...	25	WA

In [6]:

```

#sns.relplot(data=wd,x="Easting", y="Northing", hue="Employee_ID",size="Em

import plotly.express as px

import plotly.graph_objects as go

import requests, io, json

fig = px.scatter(wd, x="Easting", y="Northing", color="Working_Activity", l

fig.add_trace(go.Scatter(
    x=[8.5, 5, 2, 17.5, 17, 12, 6.5],
    y=[12.5, 4.5, 13, 17.5, 3, 3, 22],
    text=[ "C01",
           "C02",
           "C03",
           "MC01",
           "R01",
           "W01",
           "S01"],
    mode="text",
))

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=7, y0=11,
    x1=10, y1=14,
    line=dict(
        color="RoyalBlue",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=2, y0=2,
    x1=8, y1=7,
    line=dict(
        color="RoyalBlue",

```

```
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=1, y0=11,
    x1=3, y1=15,
    line=dict(
        color="RoyalBlue",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=16, y0=16,
    x1=19, y1=19,
    line=dict(
        color="RoyalBlue",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=10, y0=1,
    x1=14, y1=5,
    line=dict(
        color="RoyalBlue",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

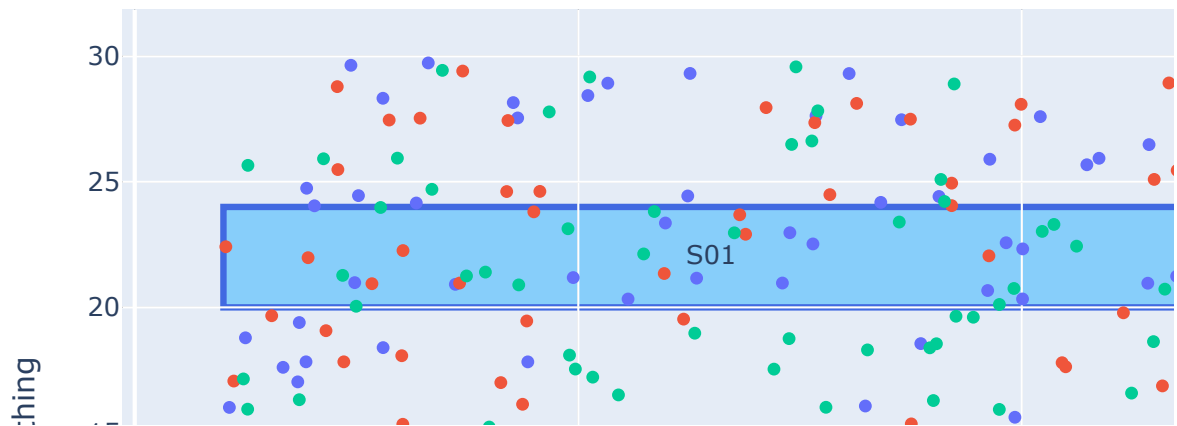
fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=1, y0=20,
    x1=12, y1=24,
    line=dict(
        color="RoyalBlue",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
    xref="x", yref="y",
    x0=16, y0=2,
    x1=18, y1=4,
    line=dict(
        color="LightSeaGreen",
        width=3,
    ),
    fillcolor="LightSkyBlue",layer = "below"
)
```

```

        fillcolor="PaleTurquoise", layer = "below"
    )
fig.show()

```



```

In [30]: #bd2.plot(figsize=(10, 10), alpha=0.55, cmap='RdBu_r')

#wd_joined.plot(figsize=(10, 10), alpha=0.55, cmap='RdBu_r')

fig = px.scatter(wd, x="Easting", y="Northing", animation_frame="Date", an:

fig.add_trace(go.Scatter(
    x=[8.5, 5, 2, 17.5, 17, 12, 6.5],
    y=[12.5, 4.5, 13, 17.5, 3, 3, 22],
    text=["C01",
          "C02",
          "C03",
          "MC01",
          "R01",

```

```
        "W01",
        "S01"],
    mode="text",
))

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=7, y0=11,
              x1=10, y1=14,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=2, y0=2,
              x1=8, y1=7,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=1, y0=11,
              x1=3, y1=15,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=16, y0=16,
              x1=19, y1=19,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=10, y0=1,
              x1=14, y1=5,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
```

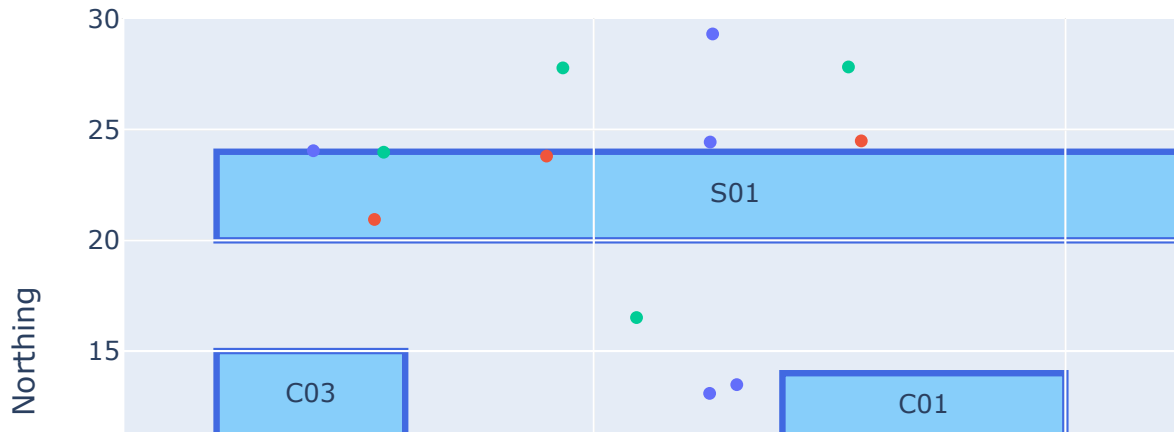


```
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=1, y0=20,
              x1=12, y1=24,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ),
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=16, y0=2,
              x1=18, y1=4,
              line=dict(
                  color="LightSeaGreen",
                  width=3,
              ),
              fillcolor="PaleTurquoise", layer = "below"
)

fig.show()
fig.write_html("path/to/file.html")
```



```
In [61]: #bd2.plot(figsize=(10, 10), alpha=0.55, cmap='RdBu_r')

import plotly.express as px

fig = px.scatter(wd, x="Easting", y="Northing", animation_frame="Date", and

trace = go.Scatter(
    x=[8.5, 5, 2, 17.5, 17, 12, 6.5],
    y=[12.5, 4.5, 13, 17.5, 3, 3, 22],
    text=["C01",
          "C02",
          "C03",
          "MC01",
          "R01",
          "W01",
          "S01"],
    mode="text",
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=7, y0=11,
              x1=10, y1=14,
```

```
        line=dict(
            color="RoyalBlue",
            width=3,
        ),col="all",row = "all",
        fillcolor="LightSkyBlue",layer = "below"
    )
    fig.add_shape(type="rect",
        xref="x", yref="y",
        x0=2, y0=2,
        x1=8, y1=7,
        line=dict(
            color="RoyalBlue",
            width=3,
        ),col="all",row = "all",
        fillcolor="LightSkyBlue",layer = "below"
    )

    fig.add_shape(type="rect",
        xref="x", yref="y",
        x0=1, y0=11,
        x1=3, y1=15,
        line=dict(
            color="RoyalBlue",
            width=3,
        ),col="all",row = "all",
        fillcolor="LightSkyBlue",layer = "below"
    )

    fig.add_shape(type="rect",
        xref="x", yref="y",
        x0=16, y0=16,
        x1=19, y1=19,
        line=dict(
            color="RoyalBlue",
            width=3,
        ),col="all",row = "all",
        fillcolor="LightSkyBlue",layer = "below"
    )

    fig.add_shape(type="rect",
        xref="x", yref="y",
        x0=10, y0=1,
        x1=14, y1=5,
        line=dict(
            color="RoyalBlue",
            width=3,
        ),col="all",row = "all",
        fillcolor="LightSkyBlue",layer = "below"
    )

    fig.add_shape(type="rect",
        xref="x", yref="y",
        x0=1, y0=20,
        x1=12, y1=24,
        line=dict(
            color="RoyalBlue",
            width=3,
```

```

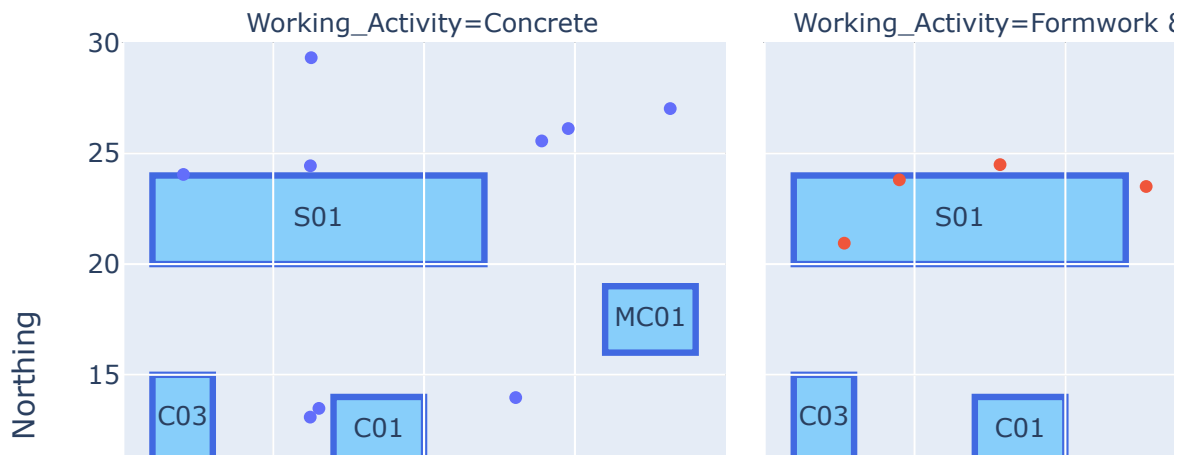
),col="all",row="all",
fillcolor="LightSkyBlue",layer="below"
)

fig.add_shape(type="rect",
xref="x",yref="y",
x0=16,y0=2,
x1=18,y1=4,
line=dict(
color="LightSeaGreen",
width=3,
),col="all",row="all",
fillcolor="PaleTurquoise",layer="below"
)

fig.add_trace(trace,row="all",col="all",exclude_empty_subplots=True)

fig.show()

```



```

In [59]: import plotly.express as px

fig = px.line(wd_joined, x="Easting", y="Northing", animation_frame="Date"

```

```

trace = go.Scatter(
    x=[8.5, 5, 2, 17.5, 17, 12, 6.5],
    y=[12.5, 4.5, 13, 17.5, 3, 3, 22],
    text=["C01",
          "C02",
          "C03",
          "MC01",
          "R01",
          "W01",
          "S01"],
    mode="text",
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=7, y0=11,
              x1=10, y1=14,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ), col="all", row = "all",
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=2, y0=2,
              x1=8, y1=7,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ), col="all", row = "all",
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=1, y0=11,
              x1=3, y1=15,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ), col="all", row = "all",
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",
              xref="x", yref="y",
              x0=16, y0=16,
              x1=19, y1=19,
              line=dict(
                  color="RoyalBlue",
                  width=3,
              ), col="all", row = "all",
              fillcolor="LightSkyBlue", layer = "below"
)

fig.add_shape(type="rect",

```

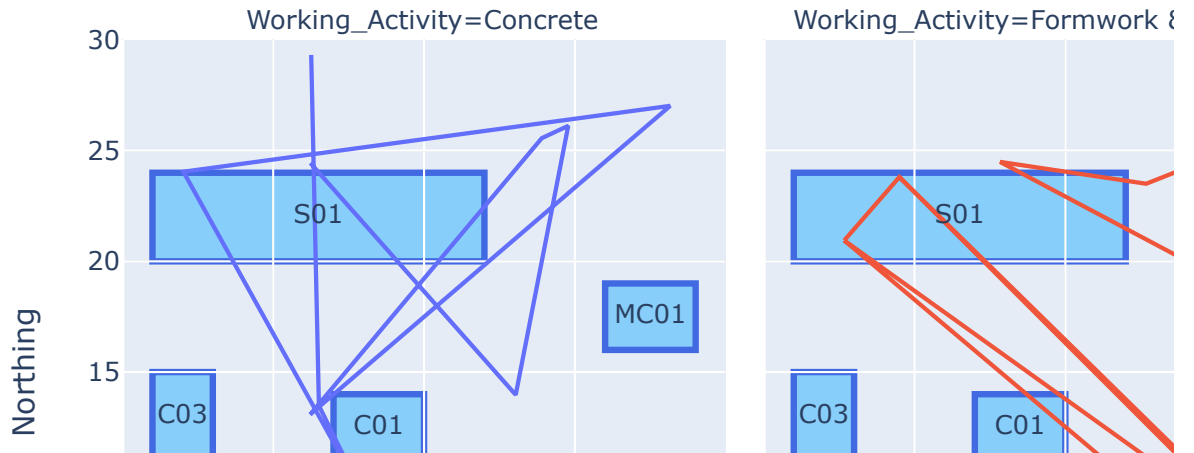
```
xref="x", yref="y",
x0=10, y0=1,
x1=14, y1=5,
line=dict(
    color="RoyalBlue",
    width=3,
),col="all",row = "all",
fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
xref="x", yref="y",
x0=1, y0=20,
x1=12, y1=24,
line=dict(
    color="RoyalBlue",
    width=3,
),col="all",row = "all",
fillcolor="LightSkyBlue",layer = "below"
)

fig.add_shape(type="rect",
xref="x", yref="y",
x0=16, y0=2,
x1=18, y1=4,
line=dict(
    color="LightSeaGreen",
    width=3,
),col="all",row = "all",
fillcolor="PaleTurquoise",layer = "below"
)

fig.add_trace(trace, row="all", col="all", exclude_empty_subplots=True)

fig.show()
```

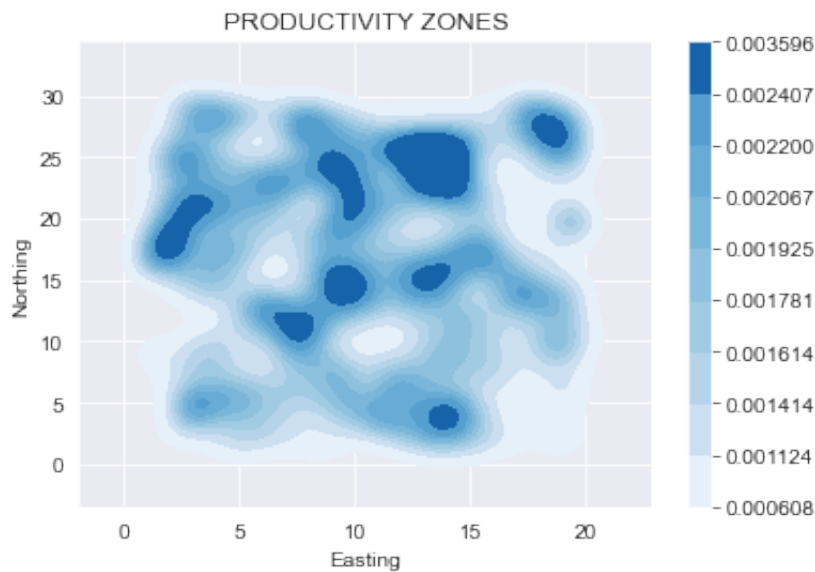


```
In [11]: Dist_Code = wd_joined['Area_Code'].value_counts(normalize=True)
Dist_Category = wd_joined['Category'].value_counts(normalize=True)
Dist_Type = wd_joined['Employee_Type'].value_counts(normalize=True)

print(Dist_Code)
print(Dist_Category)
print(Dist_Type)
```

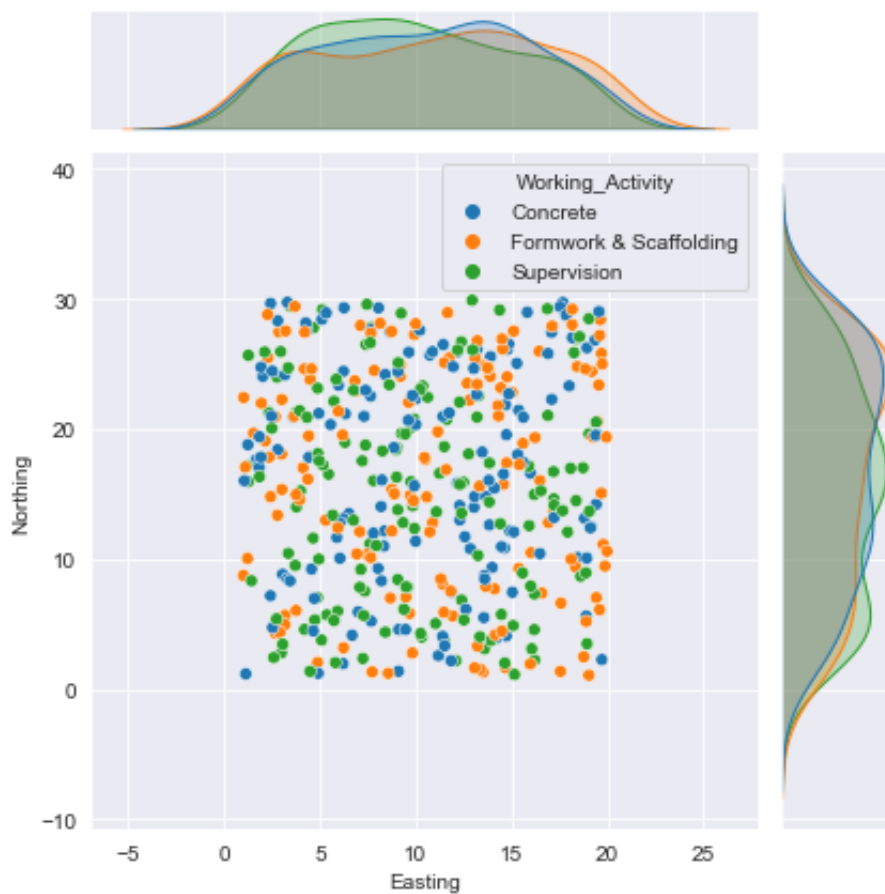
```
B      0.816697
S01     0.076225
C02     0.047187
W01     0.027223
C01     0.018149
MC01    0.007260
C03     0.003630
R01     0.003630
Name: Area_Code, dtype: float64
TA      0.816697
WA      0.179673
RA      0.003630
Name: Category, dtype: float64
Skilled Labor    0.642468
Junior Engineer  0.357532
Name: Employee_Type, dtype: float64
```

```
In [62]: sns.kdeplot(x=wd.Easting, y=wd.Northing, cbar = True, cmap="Blues", shade=True,
plt.show())
```



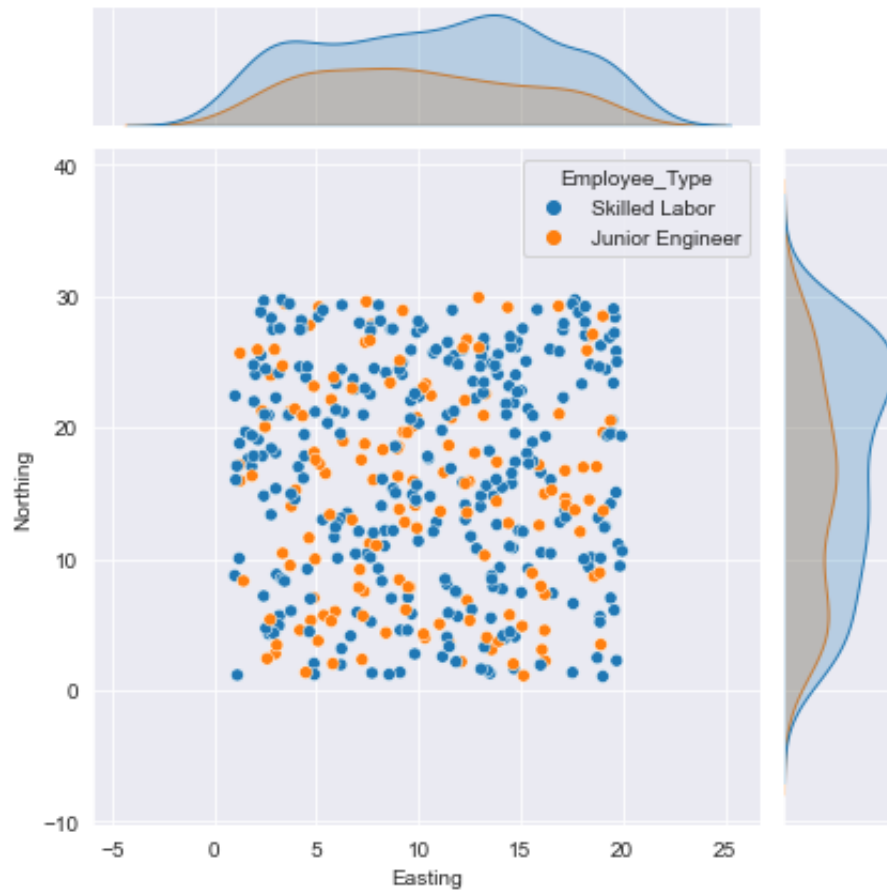
```
In [13]: sns.jointplot(data=wd, x="Easting", y="Northing", hue="Working_Activity")
```

```
Out[13]: <seaborn.axisgrid.JointGrid at 0x124f76100>
```



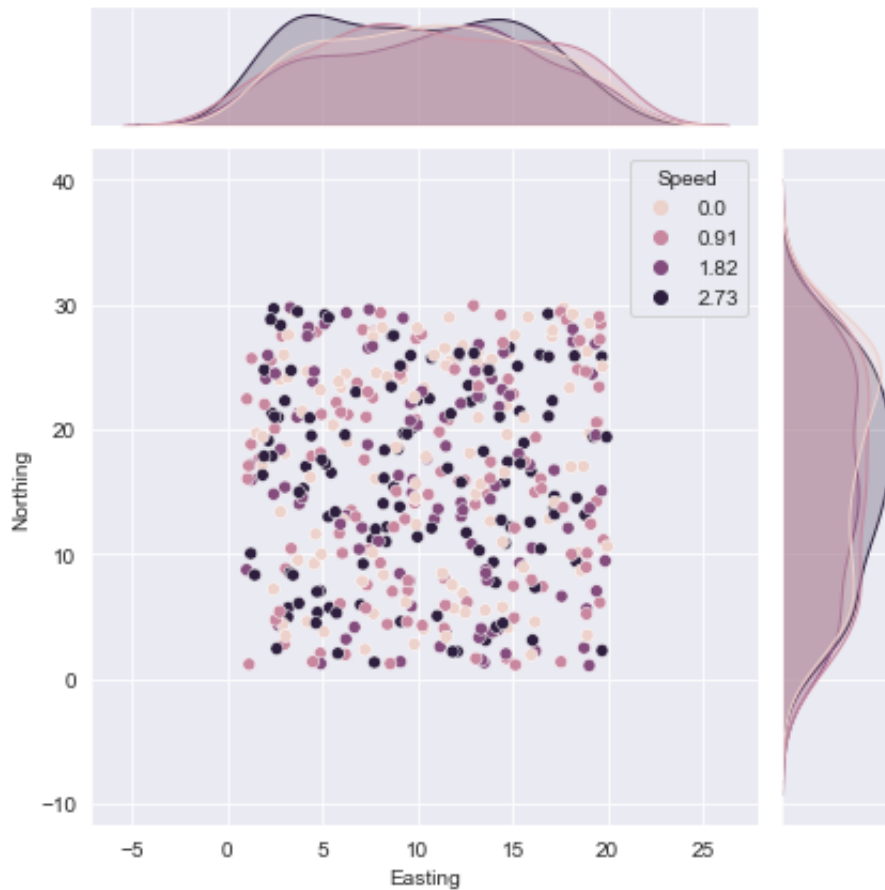
```
In [14]: sns.jointplot(data=wd, x="Easting", y="Northing", hue="Employee_Type")
```


Out[14]: <seaborn.axisgrid.JointGrid at 0x124367070>



```
In [15]: sns.jointplot(data=wd, x="Easting", y="Northing", hue="Speed")
```

Out[15]: <seaborn.axisgrid.JointGrid at 0x122e23a60>



```
In [16]: wd['xy'] = wd.apply(lambda x: [x['Easting'], x['Northing']], axis=1)

gb = wd.groupby('Employee_ID')
wd2 = gb.agg({'xy': lambda x: list(x)})

wd2

Points = wd[["Easting", "Northing"]].to_numpy()
Points

import libpysal as ps
import numpy as np
from pointpats import PointPattern

p1 = PointPattern(Points)
p1.mbb

p1.summary()
```

```
Point Pattern
450 points
Bounding rectangle [(1.013,1.052), (19.968,29.899)]
Area of window: 546.794885
Intensity estimate for window: 0.8229777057991315
      x      y
0   6.260  29.319
1   4.499  23.805
2  18.384  26.936
3   6.516  13.484
4  15.371   9.268
```

```
In [17]: from pointpats.centrography import hull, mbr, mean_center, weighted_mean_center
```

```
In [18]: mc = mean_center(p1.points)
mc

p1.plot()
plt.plot(mc[0], mc[1], 'g^', label='Mean Center')
plt.legend(numpoints=1)

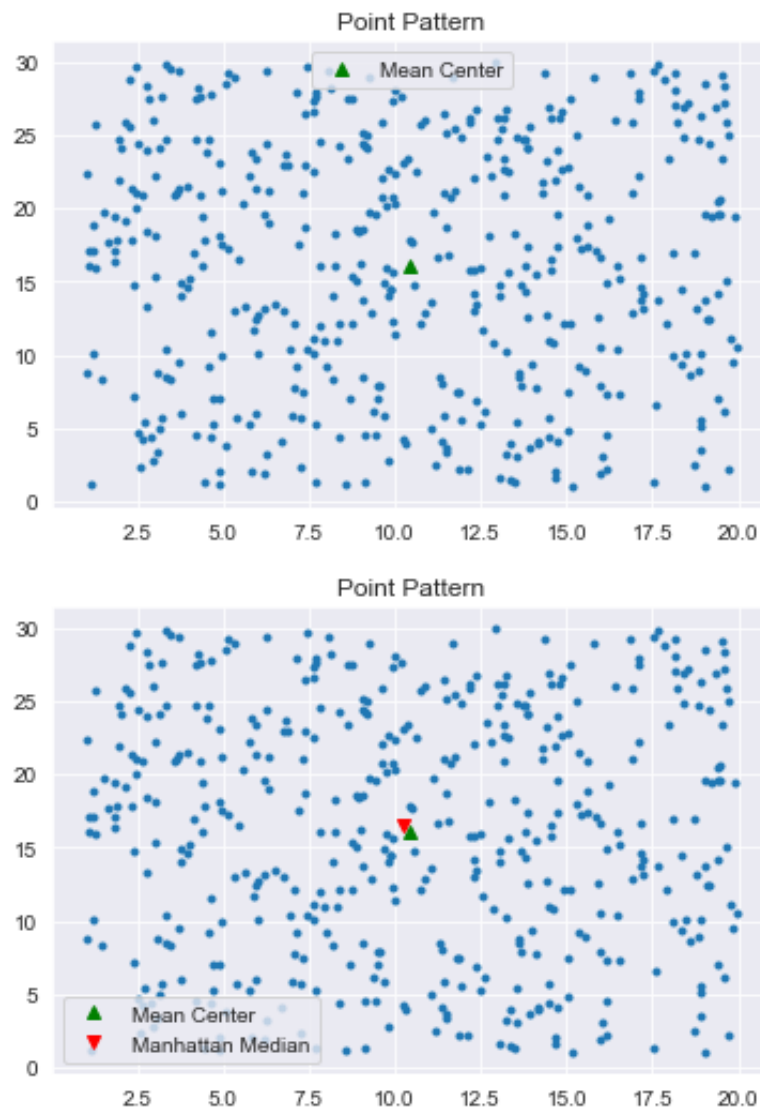
mm = manhattan_median(p1.points)
mm

p1.plot()
plt.plot(mc[0], mc[1], 'g^', label='Mean Center')
plt.plot(mm[0], mm[1], 'rv', label='Manhattan Median')
plt.legend(numpoints=1)
```

```
/opt/anaconda3/lib/python3.8/site-packages/pointpats/centrography.py:208: UserWarning:
```

Manhattan Median is not unique for even point patterns.

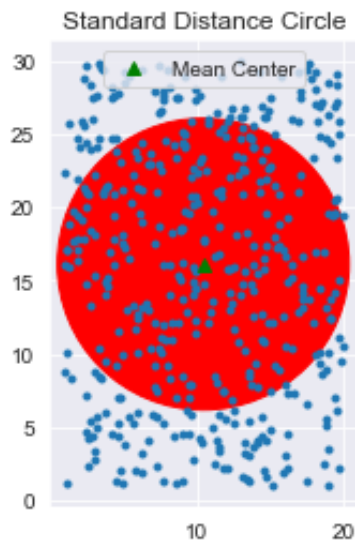
Out[18]: <matplotlib.legend.Legend at 0x12c879820>



```
In [19]: stdd = std_distance(p1.points)
stdd

circle1=plt.Circle((mc[0], mc[1]),stdd,color='r')
ax = p1.plot(get_ax=True, title='Standard Distance Circle')
ax.add_artist(circle1)
plt.plot(mc[0], mc[1], 'g^', label='Mean Center')
ax.set_aspect('equal')
plt.legend(numpoints=1)
```

Out[19]: <matplotlib.legend.Legend at 0x12c9f2790>



```
In [20]: stdd = std_distance(p1.points)
stdd

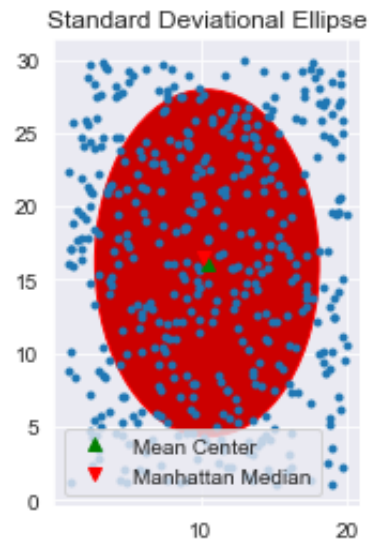
from matplotlib.patches import Ellipse
from pylab import figure, show, rand
fig = figure()

sx, sy, theta = ellipse(p1.points)
sx, sy, theta

theta_degree = np.degrees(theta) #need degree of rotation to plot the ellipse
theta_degree

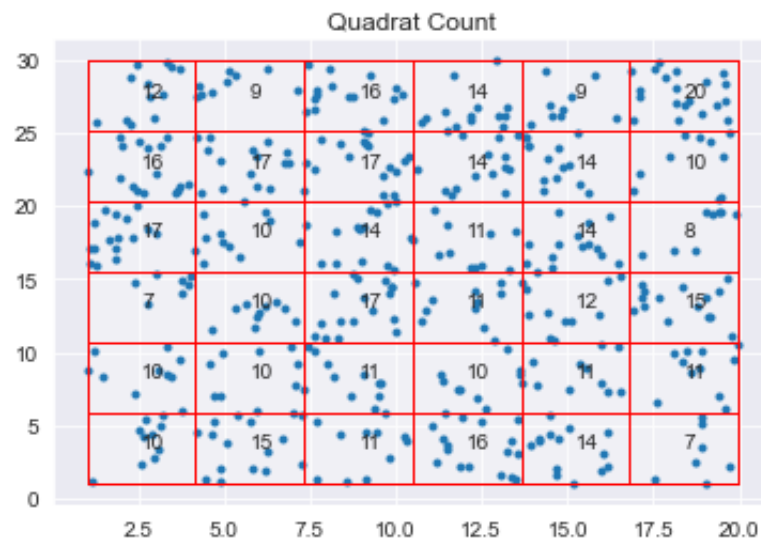
e = Ellipse(xy=mean_center(p1.points), width=sx*2, height=sy*2, angle=-theta_degree)
ax = p1.plot(get_ax=True, title='Standard Deviational Ellipse')
ax.add_artist(e)
e.set_clip_box(ax.bbox)
e.set_facecolor([0.8,0,0])
e.set_edgecolor([1,0,0])
ax.set_aspect('equal')
plt.plot(mc[0], mc[1], 'g^', label='Mean Center')
plt.plot(mm[0], mm[1], 'rv', label='Manhattan Median')
plt.legend(numpoints=1)
plt.legend(numpoints=1)
show()
```

<Figure size 432x288 with 0 Axes>



```
In [21]: from pointpats import PointPattern, as_window
from pointpats import PoissonPointProcess as csr
import pointpats.quadrat_statistics as qs

q_r = qs.QStatistic(p1, shape= "rectangle", nx =6, ny =6)
q_r.plot()
q_r.chi2
q_r.df
q_r.chi2_pvalue
```



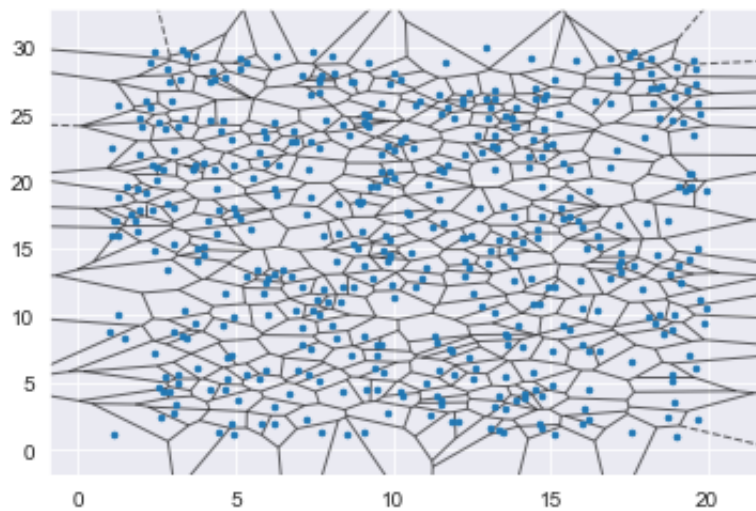
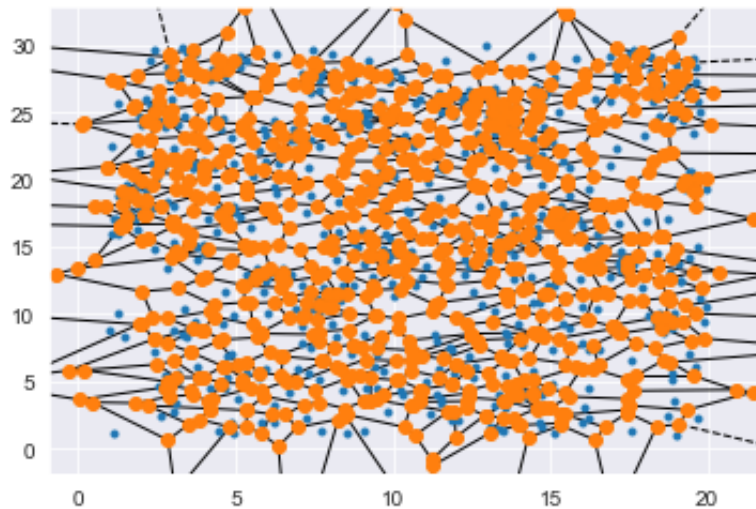
Out[21]: 0.7505998180180904

```
In [22]: import matplotlib as mpl

from scipy.spatial import Voronoi, voronoi_plot_2d
vor = Voronoi(Points)

import matplotlib.pyplot as plt
fig = voronoi_plot_2d(vor)
plt.show()

fig = voronoi_plot_2d(vor, show_vertices=False, line_colors='black',
line_width=1, line_alpha=0.6, point_size=5)
plt.show()
```



```
In [23]: import matplotlib.pyplot as plt
from sklearn.datasets.samples_generator import make_blobs
from sklearn.cluster import Birch

# Creating the BIRCH clustering model
model = Birch(branching_factor = 50, n_clusters = 12, threshold = 1.5)

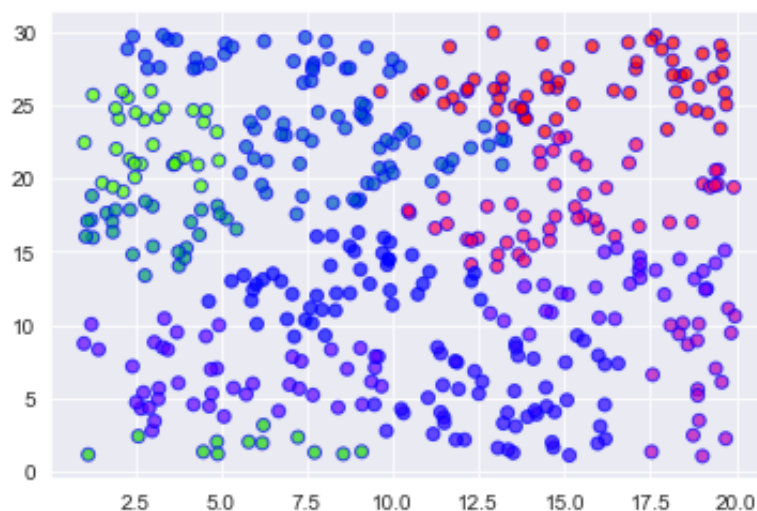
# Fit the data (Training)
model.fit(Points)

# Predict the same data
pred = model.predict(Points)

# Creating a scatter plot
plt.scatter(Points[:, 0], Points[:, 1], c = pred, cmap = 'prism', alpha = 0.5)
plt.show()
```

/opt/anaconda3/lib/python3.8/site-packages/sklearn/utils/deprecation.py:143
: FutureWarning:

The sklearn.datasets.samples_generator module is deprecated in version 0.22 and will be removed in version 0.24. The corresponding classes / functions should instead be imported from sklearn.datasets. Anything that cannot be imported from sklearn.datasets is now part of the private API.



```
In [24]: import scipy.spatial
import libpysal as ps
import numpy as np
from pointpats import ripley, g, f, k, j, l
%matplotlib inline
import matplotlib.pyplot as plt

p1.knn()
p1.knn(2)
p1.max_nnd
p1.min_nnd
p1.mean_nnd
p1.nnd
p1.nnd.sum()/p1.n
```

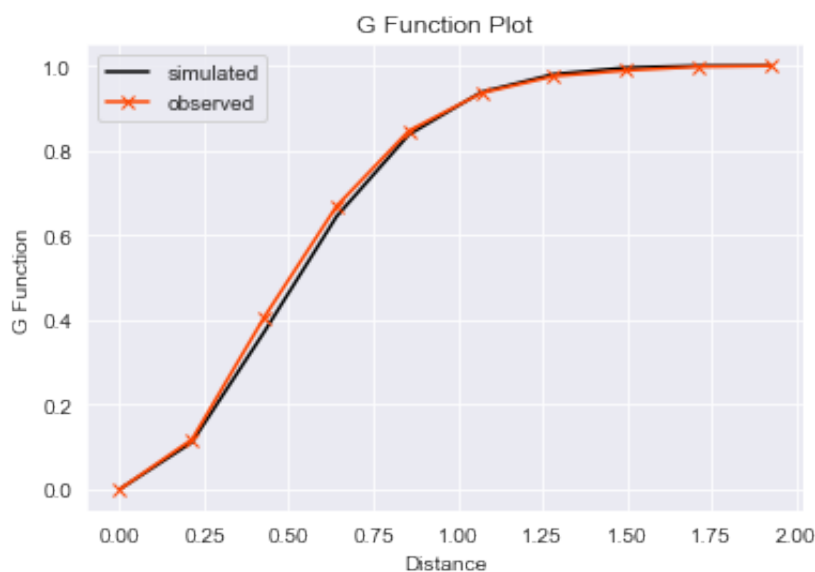

Out[24]: 0.5477341819743731

```
In [25]: from scipy import spatial
import libpysal as ps
import numpy as np
from pointpats import ripley
%matplotlib inline
import matplotlib.pyplot as plt

g_test = ripley.g_test(Points, support=10)
g_test.support
g_test.statistic
g_test.pvalue
g_test.simulations
g_test = ripley.g_test(Points, support=10, keep_simulations=True)

plt.plot(g_test.support, np.median(g_test.simulations, axis=0),
         color='k', label='simulated')
plt.plot(g_test.support, g_test.statistic,
         marker='x', color='orangered', label='observed')
plt.legend()
plt.xlabel('Distance')
plt.ylabel('G Function')
plt.title('G Function Plot')
plt.show()

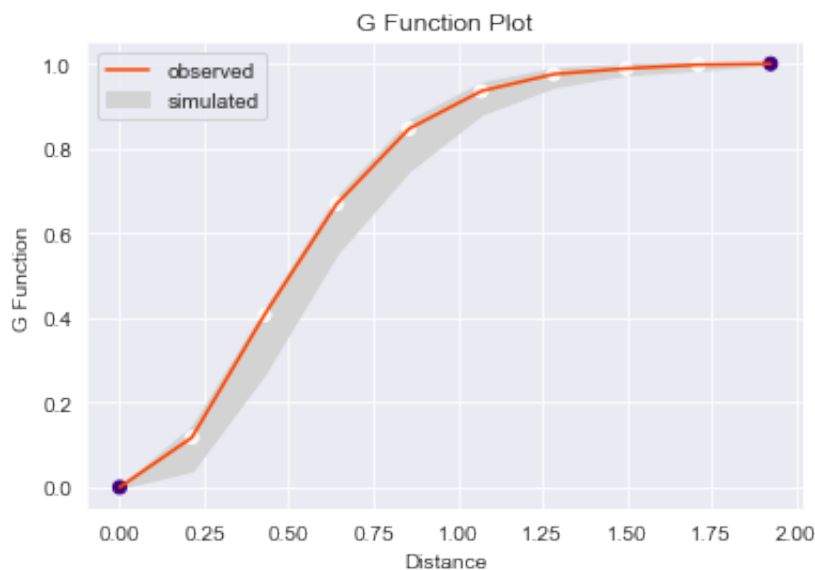
g_test.pvalue
```



Out[25]: array([0. , 0.3818, 0.118 , 0.1808, 0.3161, 0.3776, 0.2302, 0.0686,
0.2073, 0.])

```
In [26]: # grab the middle 95% of simulations using numpy:
middle_95pct = np.percentile(g_test.simulations, q=(0, 95.0), axis=0)
# use the fill_between function to color between the 2.5% and 95% envelope
plt.fill_between(g_test.support, *middle_95pct,
                 color='lightgrey', label='simulated')

# plot the line for the observed value of G(d)
plt.plot(g_test.support, g_test.statistic,
         color='orangered', label='observed')
# and plot the support points depending on whether their p-value is smaller
plt.scatter(g_test.support, g_test.statistic,
           cmap='Purples', c=g_test.pvalue < .01)
plt.legend()
plt.xlabel('Distance')
plt.ylabel('G Function')
plt.title('G Function Plot')
plt.show()
```



```
In [27]: riple.y.g_function(Points)
```

```
Out[27]: (array([0.          , 0.10129516, 0.20259032, 0.30388548, 0.40518063,
0.50647579, 0.60777095, 0.70906611, 0.81036127, 0.91165643,
1.01295159, 1.11424674, 1.2155419 , 1.31683706, 1.41813222,
1.51942738, 1.62072254, 1.72201769, 1.82331285, 1.92460801]),
array([0.          , 0.03111111, 0.11555556, 0.21111111, 0.36444444,
0.52          , 0.64888889, 0.72888889, 0.82222222, 0.88222222,
0.92          , 0.95111111, 0.96222222, 0.97555556, 0.98444444,
0.99333333, 0.99777778, 0.99777778, 0.99777778, 1.          ]))
```