STATISTICAL METHODS FOR
PREDICTING INDIVIDUAL
MOVEMENT PATTERNS

by

Jenifer R. McClary
A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Mathematics and Statistics - Statistics Specialty).

Golden, Colorado
Date ______________________

Signed: ______________________
   Jenifer R. McClary

Signed: ______________________
   Dr. William C. Navidi
   Thesis Advisor

Golden, Colorado
Date ______________________

Signed: ______________________
   Dr. Gregory E. Fasshauer
   Professor and Head
   Department of Applied Mathematics and Statistics
ABSTRACT

One of the most potentially useful methods for truly understanding our habits is by observing the data captured by our cellular phones. In 2009, Nokia launched the Lausanne Data Collection Campaign, which was designed to collect behavioral data from 170 volunteers through data collection software on contributed phones. In particular, GPS and WiFi connection data depicted the location of volunteers for the duration of their participation in the study. Through statistical analysis of the past location data for an individual over a prescribed duration of time, one can potentially predict the individuals future movement with desirable degrees of confidence. This research investigates a methodology for conducting predictive analysis of individual movement based on past known movements through analysis of several known time series statistical methods. Based on this analysis, I provide recommendations for application of this method to other datasets.
# TABLE OF CONTENTS

ABSTRACT ......................... iii
LIST OF FIGURES ..................... vi
LIST OF TABLES ...................... viii
LIST OF ABBREVIATIONS ................. ix
DEDICATION ........................ x
CHAPTER 1 INTRODUCTION .............. 1
CHAPTER 2 LITERATURE REVIEW ......... 4
CHAPTER 3 METHODOLOGY ............... 6
  3.1 Data Processing .................. 6
  3.2 Introduction to Time Series ........ 7
  3.3 Time Series Models ............... 10
  3.4 Modeling of the LDCC Time Series Dataset ........ 11
  3.5 Prediction of the LDCC Time Series Dataset ........ 13
CHAPTER 4 RESULTS ................... 14
CHAPTER 5 SUMMARY AND CONCLUSIONS .... 29
REFERENCES CITED ................... 31
APPENDIX A REMAINING USERS DATA PROCESSING ........ 33
  A.1 User 5479 ...................... 34
  A.2 User 5578 ...................... 36
  A.3 User 5927 ...................... 38
LIST OF FIGURES

Figure 4.1  Original versus Truncated versus Final Processed Data for User 5542. . . 15
Figure 4.2  Original versus Truncated versus Final Processed Data for User 5975. . . 16
Figure 4.3  Original vs Interpolated 60s vs Interpolated 60m User 5542. . . . . . . . 17
Figure 4.4  Original vs Interpolated 60s vs Interpolated 60m User 5975. . . . . . . . 18
Figure 4.5  Additivity Model User 5542. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
Figure 4.6  Additivity Model User 5975. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
Figure 4.7  Autocorrelation functions for Users 5542 and 5975. . . . . . . . . . . . . . 21
Figure 4.8  True versus Predicted Latitude and Longitude Locations, User 5542. . . 22
Figure 4.9  True versus Predicted Latitude and Longitude Locations, User 5975. . . 23
Figure 4.10 Final True versus Predicted Points User 5542. . . . . . . . . . . . . . . . . . . 24
Figure 4.11 Final True versus Predicted Points User 5975. . . . . . . . . . . . . . . . . . . 25
Figure A.1  Original versus Final Processed Data for User 5479. . . . . . . . . . . . . . 34
Figure A.2  Original vs Interpolated 60s vs Interpolated 60m User 5479. . . . . . . . . 35
Figure A.3  Original versus Truncated versus Final Processed Data for User 5578. . . 36
Figure A.4  Original vs Interpolated 60s vs Interpolated 60m User 5578. . . . . . . . . 37
Figure A.5  Original versus Truncated versus Final Processed Data for User 5927. . . 38
Figure A.6  Original vs Interpolated 60s vs Interpolated 60m User 5927. . . . . . . . . 39
Figure A.7  Original versus Truncated versus Final Processed Data for User 5973. . . 40
Figure A.8  Original vs Interpolated 60s vs Interpolated 60m User 5973. . . . . . . . . 41
Figure B.1  Additivity Model User 5479. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
Figure B.2  Autocorrelation function for User 5479. .......................... 43
Figure B.3  Additivity Model User 5578. ........................................ 44
Figure B.4  Autocorrelation function for User 5578. .......................... 45
Figure B.5  Additivity Model User 5927. ........................................ 46
Figure B.6  Autocorrelation function for User 5927. .......................... 47
Figure B.7  Additivity Model User 5973. ........................................ 48
Figure B.8  Autocorrelation function for User 5973. .......................... 49
Figure C.1  True versus Predicted Latitude and Longitude Locations, User 5479. . . . 51
Figure C.2  Final True versus Predicted Points User 5479. .................... 52
Figure C.3  True versus Predicted Latitude and Longitude Locations, User 5578. . . . 54
Figure C.4  Final True versus Predicted Points User 5578. .................... 55
Figure C.5  True versus Predicted Latitude and Longitude Locations, User 5927. . . . 57
Figure C.6  Final True versus Predicted Points User 5927. .................... 58
Figure C.7  True versus Predicted Latitude and Longitude Locations, User 5973. . . . 60
Figure C.8  Final True versus Predicted Points User 5973. .................... 61
LIST OF TABLES

Table 4.1 Individual User Dataset Length ................................................. 14

Table 4.2 User 5542 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 27

Table 4.3 User 5975 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 27

Table 4.4 Mean MSE and Prediction Accuracy by Results Group .................. 28

Table A.1 Individual User Dataset Length for Remaining Users .................. 33

Table C.1 User 5479 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 50

Table C.2 User 5578 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 53

Table C.3 User 5927 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 56

Table C.4 User 5973 Data Models by Input Percentage, MSE (Lat/Long), Total
MSE, and Resulting Prediction Accuracy .................................................. 59
LIST OF ABBREVIATIONS

Lausanne Data Collection Campaign . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . LDCC
Global Positioning System . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . GPS
Wireless Local Area Network . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . WLAN
Mobile Data Challenge . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . MDC
Autocorrelation Function . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ACF
Moving Average . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . MA
Autoregressive . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . AR
Autoregressive Moving Average . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ARMA
Autoregressive Integrated Moving Average . . . . . . . . . . . . . . . . . . . . . . ARIMA
Seasonal Autoregressive Integrated Moving Average . . . . . . . . . . . . . . . . SARIMA
Akaike Information Criteria . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . AIC
Bayesian Information Criteria . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . BIC
Mean Squared Error . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . MSE
I lovingly dedicate the time and effort committed to this research to my daughters,

Madison and Finley.

“Always remember, you have within you the strength, the patience, and the passion to reach for the stars to change the world.” -Harriet Tubman
CHAPTER 1
INTRODUCTION

Consider an individual and his daily life. He wakes up, prepares for his day, commutes to his source of income, performs tasks throughout the day to earn an income, commutes to other necessary life destinations such as the grocery store or gym, and returns to his residence. Most days of the week, this routine is replicated with few changes, whereas a few other days of the week, he may have drastic differences. Many of us can relate to a similar routine and expect that if someone wanted to anticipate our location, it shouldn’t be much of a challenge.

Reconsider that same individual but now with more specifics: he is a criminal. When he departs his residence, he is committing crimes at unanticipated locations though still engaging in necessary life activities. By predicting the route for which he takes to commit a criminal activity or the specific location of the crime, we may be able to intercept the event altogether thus preventing future events. By studying an individual’s location life patterns over time, we may predict future locations with increasing precision.

In 2009, Nokia launched the Lausanne Data Collection Campaign (LDCC) followed by the Mobile Data Challenge (MDC). For the campaign, 170 volunteers from the Lake Geneva region of Switzerland received Nokia N95 phones with non-intrusive data collection software. Users were frequently networked through work or social connections and were recruited using a technique where initial volunteers were encouraged to invite friends and family members to volunteer, incorporating a natural social network in the campaign. Volunteers participated over a duration of seventeen months with a large variance of activity by individual from a few days to many months [1].

The software installed on volunteers’ phones collected system information, social communication, user agenda and contacts, phone usage, networking data, and location and...
movement data. All collected information is paired with initial user form information which includes an assigned user identification code, demographics, location labels, and assigned device International Mobile Equipment Identity (IMEI). For the purpose of this study, I chose to restrict inputs to location and movement data which consisted of Global Positioning System (GPS) data, Wireless Local Area Network (WLAN) data and acceleration data [1].

The methodology by which the location data is collected varies by the “state” of the device. When the device is connected to a known WLAN, the device location is recorded every 120 seconds. If connected to a known WLAN and the acceleration of the device exceeds a prescribed threshold, the location is recorded every 60 seconds. When the phone is stationary and connected to a known WLAN, location is recorded every 600 seconds. When outdoor and mobile, the device records GPS location information continuously. When battery levels are minimal, location data is a lesser priority; GPS location is not recorded and WLAN is recorded every 900 seconds [1].

The purpose of this study is to receive inputs of various duration of individuals’ location information, assess for patterns, and perform predictive analysis with defined levels of confidence. The LDCC dataset is a resource with location data for individuals for time periods varying from a couple of days to several months in duration. By observing an initial portion of an individual’s location data, several statistical methods (e.g. spatio-temporal) can be implemented to detect and model trend and seasonality to predict the individual’s location data in the latter portion of the dataset. For example, if an individual’s location data covers a time period of three weeks, we can observe the first two weeks of data, model the trend and seasonality, and predict the final week of the data.

By implementing multiple existing statistical methods for detection and modeling of trend and seasonality, we can assess the methods for their accuracy with regards to the individuals in this dataset. By comparing the predicted data to the true data, we can develop a confidence interval for each of the methods per individual and determine the level
This study aims to answer the following question: Given a prescribed input duration of location data for an individual, with what degree of confidence can we predict future location patterns for the individual?
CHAPTER 2
LITERATURE REVIEW

As part of the LDCC, Nokia announced the Mobile Data Challenge, an international data analysis challenge for non-profit organizations and academics to conduct analysis and present unique conclusions regarding limited data from the LDCC. Participants opted to conduct research in one of several categories. Among the categories, one focused on individual and group locations with the potential to forecast future movements [1]. As a result, literature regarding location and predictive analysis of the LDCC dataset is vast, though all taking differing perspectives from this report. Therefore, resources are plentiful for varying analytic perspectives on predictive analysis for this dataset.

Most comparable to my research perspective is the work of Domenico et al, who assessed multivariate, non-linear time series prediction techniques. Domenico et al predicted an individual’s future movement pattern by considering the movement of others who had correlated mobility patterns. Therefore, they used the trends of individuals who previously completed similar mobility patterns and predicted an individual’s future mobility from those of others who completed similar paths. By conducting this analysis, they were able to increase forecasting accuracy. Domenico et al also evaluated their methodology against Autoregressive Moving Average (ARMA) models, which they considered to be sub par to the methodology that they employed [2].

A trend within the research is the categorization of locations. Within the LDCC dataset, many users provide labels of locations to include home, “work,” “sport,” etc. Soikkeli established a “semantic place inference” that used the labels provided and implemented an algorithm to infer other labels based on those provided. He assessed the life patterns using Naive Bayes, Bayes Networks and a decision tree algorithm to predict future locations. Soikkeli achieved approximately 77% accuracy in his model with only 1 week of data and
approximately 85% accuracy given 6 weeks of prior data [3].

Given a requirement for a “light” algorithm and computation designed for implementation from a smartphone, Do et al designed a tailored interface for an individual user. The interface was based on where an individual was assessed to travel next by using a kernel density algorithm. Because of their desire for a “light” design, they did not use a central database of known locations and only accessed limited previously travelled locations of the individual. Though their research is aligned with the goals of this report, their future prediction was focused on very short-term windows (within 10 minutes), versus predicting within upcoming hours and days [4].

Other analysis techniques employed included spatio-temporal clustering [5], Markov Models [6], Monte Carlo Simulations [7], and machine learning methods [8].

Results from a specific work that I am interested in incorporating in my analysis is that of Etter et al, for which a significant result was the concept of “aging” data. “Aged” data refers to the value of recent habits over the trend of the accumulated dataset. Specifically, for analyzing the life patterns of an individual, recent past patterns hold greater value than the patterns that best fit the entire dataset [9].
CHAPTER 3
METHODOLOGY

The LDCC dataset was obtained by request from the Idiap Research Institute in November 2018. Though the entire dataset was provided, the “Location and Movement” and “Records” categories were the only portions of the dataset considered for the analysis of this report. Within the “Location and Movement” data, assessed data was restricted to the category called “GPS” [See Appendix D for “MDC Database Organizational Chart”].

3.1 Data Processing

Upon initial observation of the data, it was evident that extensive data formatting would be necessary. Within the GPS data file, several relevant categories are represented to include database key (reference number for prescribed user), time, latitude, and longitude. Since all of the data was represented in one column, it was necessary for analysis purposes to separate each of the necessary values into their own columns. This proved challenging because the database key and time were consistent lengths of characters while the accuracy recorded for latitude and longitude varied from $10^{-2}$ to $10^{-14}$. Due to the arduous process of formatting the datasets, I chose to focus my efforts on six individuals as case studies. Initially, the six individuals were chosen at random. However, initial analysis indicated that these individuals’ GPS data was too short in duration for application purposes. Therefore, I determined the six individuals with the largest GPS files and proceeded with the analysis.

Once each of the individuals’ location data was plotted, it was observed that the GPS data was concentrated over a shorter subset of the full data. For example, if an individual’s full GPS dataset covered a duration of four weeks, the individual may have only had consistent GPS recordings over a concentrated two weeks of the time period with a few sporadic occurrences, extending the recording to four weeks. Based on the problem statement of this research, it is necessary that a prediction for a future location of an individual is dependent
on recent and consistent known locations. Therefore, predicted location is not to be obtained when the location of an individual is unknown for a significant portion of time. Based on this assumption, I chose to truncate the data to the concentrated portion of GPS recordings.

Noise in the location data was also evident in the initial data analysis. For example, when plotting an individual’s longitude over time, some recordings suggested a significant change in longitudinal location over short periods of time with immediate return to the original location. I assessed these recordings as “noise” and identified such recordings by a simple velocity formula. If the velocity exceeded 160 kilometers per hour, I chose to categorize the points as noise and omit the observations. In calculating for velocity, I found that a significant number of points for each individual had multiple location recordings for a single time period, all of which were omitted.

Due to the collection methodology implemented by the LDCC, GPS-recorded locations vary in frequency depending on the “state” of the device. For example, locations may be recorded every second if the device is outdoors and mobile, but could have a lag of many hours if the battery level is low. For implementation of many time series algorithms, equal lag between occurrence of events is required. In order to create a time series dataset in adherence to this requirement, I created an Interpolated Dataset from the original by determining the individuals’ location every 60 minutes. In some instances, the location was known for that exact time period; however, in most circumstances, the location had to be interpolated based on the last known and next known position around the 60-minute time period. For this, I assumed a straight-line change in location and implemented a simple slope formula, assigning the interpolated location as the point on the line for which the 60-minute time period fell. By building an algorithm to conduct this methodology, I could easily change the frequency of measurement to any time period desired.

3.2 Introduction to Time Series

Though the LDCC dataset has many types of information, I chose a time series approach to analyze individual location data, specifically focused on the sequence in which locations are
recorded. A time series is a dataset in which events are recorded sequentially with respect to time. Typically, time series events are recorded at equal intervals (i.e. daily, monthly, weekly, annually); however, as discussed in the Data Processing section of this report, the recording frequency of this dataset was dependent on the state of the device. Therefore, I implemented an interpolation algorithm to translate the irregularly spaced location recordings to equally spaced at hourly intervals, defining an interpolated point at time \( t \) in latitude as

\[
\text{latitude}(t) = s_{t+1,t-1} \text{time}_t + b_{t+1,t-1}
\]

(3.1)

where \( s \) is the slope of the line between the next and previous point, defined as \( \frac{\text{latitude}_{t+1} - \text{latitude}_{t-1}}{\text{time}_{t+1} - \text{time}_{t-1}} \) and \( b \) is the intercept of the line between the next and previous point, defined as \((\text{latitude}_{t-1} - (\frac{\text{latitude}_{t+1} - \text{latitude}_{t-1}}{\text{time}_{t+1} - \text{time}_{t-1}}) \text{time}_{t-1})\)

To model time series data and to predict future events, there are specific data features that are necessary to identify: additivity, trend, and seasonality.

A model is additive if it is comprised of a trend \((m)\), seasonality with known period \((S)\), or both. Additivity is represented as

\[
X_t = m_t + S_t + \epsilon_t
\]

(3.2)

where \( \epsilon \) is a random error with \( E(\epsilon) = 0 \) [10].

Trend is a noteworthy attribute of many time series datasets and can be defined as a change in the average of the dataset over time. For example, if we assume a time period of one week and we have daily data, we would estimate the trend at time \( t \) as

\[
\hat{m}_t = \frac{X_{t-3} + X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2} + X_{t+3}}{7}
\]

(3.3) [10].

Next, we can assess a dataset for seasonality. We first de-trend the data by subtracting the trend from the full dataset \((D)\). Based on our estimate of the length of a cycle \((k)\) and the amount of times it has been repeated in the dataset \((p)\), we can calculate the mean cycle
at time $t$ as
\[ C_t = \frac{D_t + D_{t+k} + \ldots + D_{t+(p-1)k}}{p} \] (3.4)

The seasonality ($S$) is estimated by repeating the mean cycle $p$ times [10].

Differencing a time series is another method for removing a trend. When differencing is conducted, we remove the difference from a single event and the next sequential event. On the first iteration, this is called first differences and can be applied an unlimited amount of times, each time removing more of the trend. The first and second differences are

First Differences: $\nabla X_t = X_t - X_{t-1}$ (3.5)

Second Differences: $\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$ (3.6)

Given base knowledge of trend and seasonality, it is necessary to understand three attributes of a time series dataset prior to modeling the data: additivity, stationarity and autocorrelation.

Moreover, a time series ($X_t$) is considered weakly stationary, herein referred to as “stationary”, if

1. $E(X_t)$ is constant for all $t$,
2. $Var(X_t)$ is constant for all $t$,
3. And for any lag $h$, $Cov(X_t, X_{t+h})$ is constant for all $t$ [10].

Lastly, autocorrelation is the correlation between events in a stationary time series ($X_t$). The autocorrelation function (ACF) is defined similarly to the correlation function:

\[ \rho(h) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}} \] (3.7)

where $h$ is any non-negative integer [10].
When the ACF is plotted, autocorrelation changes over time are observed. It is expected that for random white noise processes, ACF values are high for short lags and decrease as the lag increases. However, differing behavior indicates the existence of certain types of time series processes, which is relevant to the modeling portion of this report [10].

### 3.3 Time Series Models

A Moving Average (MA) process \((X_t)\) of order \(q\) (MA\((q)\)) is a model of a random process \((Z_t)\). An MA process requires stationary data and is dependent on \(Z_t\) and \(q\)-past values of \(Z_t\). Where \(\mu, \beta_0, \beta_1, ..., \beta_q\) are real numbers and \(q\) is a positive integer, an MA\((q)\) process is defined as

\[
X_t = \mu + \beta_0 Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q}
\] (3.8)

and \(Z_t, Z_{t-1}, ..., Z_{t-q}\) are independent and identically distributed Normally with a zero mean and constant variance [10].

Next, Autoregressive Models (AR), another time series method for modeling a purely random process, are not necessarily stationary. AR Models are often identified in the ACF by periodic, wavelike behavior centered around zero with decreasing amplitude. An AR process of order \(q\) (AR\((q)\)) is defined as

\[
X_t = \alpha_1 (X_{t-1} - \mu) + ... + \alpha_q (X_{t-q} - \mu) + Z_t
\] (3.9)

where \(Z_t\) is the error term [10].

AR models resemble multiple regression models since the greater the value of \(q\), more terms are included in the model. For prediction purposes, we wish to obtain a balance of adequate fit while avoiding over-fitting of the data with excessive terms.

It is feasible to combine the attributes of AR and MA models into an Autoregressive Moving Average (ARMA) Model. ARMA models of orders \(p\) and \(q\), denoted ARMA\((p,q)\), are defined as

\[
X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q}
\] (3.10)
Furthermore, Autoregressive Integrated Moving Average (ARIMA) models incorporate the principles of AR and MA while by observing the differences of the time series data. This is applicable for datasets that are assessed to have stationary data when de-trended. We define an ARIMA model of orders \( p, d, q \), denoted ARIMA\((p,d,q)\), as

\[
W_t = \alpha_1 W_{t-1} + \ldots + \alpha_p W_{t-p} + Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q}
\]  

(3.11)

where \( d \) is a positive integer and \( W_t = \nabla^d X_t \) \( [10] \).

To note, when \( d = 0 \) in an ARIMA model, the resulting model is an ARMA\((p,q)\) model \( [10] \).

Similarly to ARIMA, Seasonal Autoregressive Integrated Moving Average (SARIMA) models include the non-seasonal values of ARIMA\((p,d,q)\) while incorporating a seasonal effect of orders \( P, D, Q, \) and \( S \). The parameters are denoted as SARIMA\((p,d,q,P,D,Q,S)\) where \( S \) is the time span of the seasonal repeating pattern. \( S \) is estimated by observing the ACF for unusual spikes or transitions in ACF values at particular lags or by information known about the collected data \( [11] \).

ARIMA and SARIMA parameters are estimated by using the Yule-Walker Equations, a default calculation in the programming language of R \( [10] \).

### 3.4 Modeling of the LDCC Time Series Dataset

In preparation for modeling each individual’s location time series dataset, I conducted a decomposition of the additive model and plotted the ACF over time. The decomposition of the additive model plots the original time series, the trend, the seasonality and the resulting error or noise. By observation, the plots can be visually assessed for stationarity. Based on the proposed problem statement, I anticipate the cycle to be 24 hours. Therefore, the autocorrelation values should be significant around the value of 24 or multiples of 24. This suggests that an individual’s schedule or routine is consistent with the past day or
several days. Observing these plots allows the confirmation or denial of the assumption of
stationarity and logical seasonality.

My strategy to model each individual’s location time series dataset was to implement
ARIMA and SARIMA Models with an input of the last 90% of the dataset, the last 70%
of the dataset, and the last 50% of the dataset. From these models, I calculated a 95%
prediction interval of the final 10% of the dataset. I compared the predicted values to the
true values to calculate a prediction accuracy.

I began by fitting ARIMA models to each individuals dataset of varying length (last
90% of total length, last 70% of total length, and last 50% of total length) by longitude
and latitude separately. The ARIMA models consisted of variations of $p = (0, 2)$, $d=(0, 2)$,
and $q=(0, 4)$. The following limitations were applied to the model: the AR portion to the
range of zero to two in order to avoid over-fitting the model for better prediction accuracy,
the differences portion to second differences since I assessed that further differencing would
return minimal returns for the computational requirement, and the MA portion of the model
to four. An assessment of each of the models was conducted by ordering by minimized Akaike
Information Criteria (AIC) and Bayesian Information Criteria (BIC), two methods for model
selection. Based on the minimized AIC and BIC models, I selected the lowest order model
between the two values if the models differed.

The ARIMA model orders of $p$, $d$, and $q$ were implemented for the values of SARIMA($p$,
$d$, $q$, $P$, $D$, $Q$, $S$) and the orders of $P$, $D$, and $Q$ were then calculated using a similar method
to before, though the values of the order of the seasonal portions were limited to $P=(0, 1)$,
$D=(0, 1)$, and $Q=(0, 2)$. The order of $S$ was derived from the results of the ACF plot, but
$S$ was consistently assumed as a value of 24. As with the ARIMA model, I calculated the
AIC and BIC values of the SARIMA models to select the lowest order model with minimized
AIC and BIC.

Each ARIMA and SARIMA model was assigned with respect to the duration of time
covered for both longitude and latitude. For example, an ARIMA and SARIMA model for
an individual for the data containing 90% of the dataset differed from the models assigned
for the data containing the 70% and 50% of the dataset. This method resulted in twelve
models for each individual: ARIMA and SARIMA for longitude, latitude, and each dataset
length (90%, 70% and 50% data).

3.5 Prediction of the LDCC Time Series Dataset

From the assigned ARIMA and SARIMA models, I conducted a prediction of 10% of
each individuals dataset by longitude and latitude. I married the predicted longitude and
latitude values to generate predicted grid points. Since the true values of the final 10% of
the dataset are known, I compared the prediction intervals of the predicted values to the
true values to calculate a prediction accuracy, defined by

\[
\frac{\text{number of true point values within prediction interval}}{\text{number of data points in 10% of individual’s dataset}}
\]

(3.12)

In order to more precisely measure accuracy between the predicted points and the true
points, I calculated the Mean Squared Error (MSE) of each of the ARIMA and SARIMA
predictions by input duration.
CHAPTER 4
RESULTS

For the results portion of this report, I will specifically focus on two individuals: Users 5542 and 5975. These individuals are representatives of two distinct results groups: (1) a group whose location data was best modeled by considering the full 90% duration of the dataset (herein referred to as the Full Data Group) and (2) a group whose location data was best modeled by only considering 50 to 70% of an individual’s data (herein referred to as the Time Sensitive Group).

In Figure 4.1 and Figure 4.2 (pages 15 and 16, respectively), I have depicted the visual transitions that occurred from the original data, to the truncated data, to the final processed data. From the original data, both users depict large gaps in reported locations. For realistic application, I implemented an assumption that I would only attempt to model location data for an individual who had a recent active location history. For this purpose, I truncated both individuals data in order to focus on a concentrated portion of frequent location data. The final processed data was obtained after removing the obscure location recordings that violated the individual being in two places at one time or moving at an unrealistic velocity (in excess of 160 kilometers per hour). As seen in Table 4.1, the Original Data for Users 5542 and 5975 is about 10% greater in length than the resulting Truncated Data and the Final Processed Data. This concept is visually evident in Figure 4.1 and Figure 4.2 in that the Original Data for both Users is highly concentrated on one end of respective time duration and once truncated, provides much more detail to location.

<table>
<thead>
<tr>
<th>Table 4.1 Individual User Dataset Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>User 5542:</td>
</tr>
<tr>
<td>User 5975:</td>
</tr>
</tbody>
</table>
Since time series algorithms require even time increments between events, I generated a dataset from the original dataset for each user that created even time increments by interpolation. I began by interpolating locations every 60 seconds. The 60-second interpolation generated more data points than the original dataset, which proved to be computationally demanding in model development. Therefore, I compared the 60-second interpolation to a 3600-second interpolation (60 minutes) and found that the general structure of the location patterns was preserved while significantly improving the computational demand. Figure 4.3 and Figure 4.4 (pages 17 and 18, respectively) demonstrate the results of the implemented interpolation method. When interpolating the location every 60 seconds, the number of points generated exceeded the number of actual points, causing greater computational demand at later points in the research. Interpolation of location points every 60 minutes generally captured the trends and changes in location while limiting the computational burden.
Figure 4.2 Original versus Truncated versus Final Processed Data for User 5975.
Figure 4.3 Original vs Interpolated 60s vs Interpolated 60m User 5542.
Figure 4.4 Original vs Interpolated 60s vs Interpolated 60m User 5975.
In order to evaluate the assumption of an additive model, I decomposed the data, estimating trend and seasonality and generated ACF plots by latitude and longitude. Figure 4.5 and Figure 4.6 (page 20) reflect the assumed additive model. User 5542’s additive model depicts a significant random error whereas User 5975’s random error is less extreme. Of significance in both random errors is that the assumption of stationarity is supported by the approximately constant variance.

Figure 4.5 Additivity Model User 5542.
Figure 4.6 Additivity Model User 5975.
The ACF plots in Figure 4.7 demonstrate high correlation at low lags which taper as lag increases. User 5975’s ACF depicts a clear wave-like pattern is evident.

Upon completion of data processing and confirmation of necessary conditions, I proceeded to modeling the data. When considering the full dataset for an individual, I implemented ARIMA and SARIMA models to model the first 90% of the dataset for prediction of the final 10%. Then, I compared this model to the model of the final 70% of the dataset and the final 50% of the dataset. Lastly, the respective predictions of the final 10% of the dataset were evaluated in relation to the first model. Results and accuracy of predictions resulted in trends aligned with two categories: Full Data and Time Sensitive Data. Users whose results aligned with the Full Data Group had decreased prediction accuracy and increased MSE as the amount of inputted data decreased. Users whose results aligned with the Time Sensitive Data had increased prediction accuracy and decreased MSE as the amount of inputted data decreased.

Figure 4.8 and Figure 4.9 (pages 22 and 23, respectively) graphically represent the true latitude and longitude values by individual as compared to the predicted values given varying
input duration. When input is maximized at 90% of the data, the predicted locations have the narrowest intervals. For User 5542, highest precision is obtained during 70% and 90% data input, while for User 5975, prediction accuracy is highest at 50% data input. As data input decreases to 50%, the prediction interval widens for both users. This is further depicted in Figure 4.10 and Figure 4.11 (pages 24 and 25, respectively) a balance and concentration of points for User 5542 is observed when data input is maximized and User 5975 when data input is minimized.

Figure 4.8 True versus Predicted Latitude and Longitude Locations, User 5542.

Table 4.2 and Table 4.3 (page 27) report the respective ARIMA and SARIMA model applied by latitude and longitude, resulting MSE, total MSE of the final predicted points, and the prediction accuracy by percent data input. Of significance, User 5542’s MSE and prediction accuracy is best at 70% data input for the ARIMA model and 90% data input for the SARIMA model. Contrastingly, User 5975’s models are both optimized with less data
input. Though User 5542 has a consistent 100% prediction accuracy, User 5975’s models resulted in greater precision and narrower prediction intervals.

When considering the average MSE and prediction accuracy by group in Table 4.4 (page 28), the Time Sensitive Group has the highest prediction accuracy with the ARIMA model with 50% data input whereas the highest prediction accuracy for the Full Data Group resulted from the ARIMA model at 90% data input. The mean results when considering all six individuals evaluated occurred with the ARIMA model at 70% data input.

Among the six individuals evaluated, the two individuals with the longest time duration in the dataset fell within the Time Sensitive Group while the four shorter duration datasets fell within the Full Data Group. The average dataset length for the Time Sensitive Group was 7.42 days compared to 5.23 days for the Full Data Group.
Figure 4.10 Final True versus Predicted Points User 5542.
Figure 4.11 Final True versus Predicted Points User 5975.
As depicted in Table 4.4 (page 28), the SARIMA models consistently performed worse than the ARIMA models with regards to both MSE and prediction accuracy. Though based on the problem statement, we expect an element of seasonality within our additive model; I assess that the time duration of the input was not sufficient for estimating a mean cycle or seasonality. Therefore, when the SARIMA model forces this estimate, the model results in a loss of prediction accuracy as compared to the ARIMA model. Of note, the Time Sensitive Group had an average of a 103% increase in MSE from ARIMA to SARIMA models while the Full Data Group had an average of 120% increase in MSE from ARIMA to SARIMA models. This potentially suggests that as the duration of the time input of the model increases, we can expect equal or greater accuracy from the SARIMA model when compared to the ARIMA model.

Additionally, Figure 4.8 and Figure 4.9 (pages 22 and 23 respectively) illustrate that the prediction accuracy of the models were a result of the size of the prediction intervals. Since the model inputs lacked sufficient duration, the model confidence interval was large, resulting in an even larger prediction interval. For example, given 90% of the data, the confidence interval for strictly the prediction in latitude for Users 5542 and 5975 was approximately 80 kilometers and 67 kilometers wide, respectively. Therefore, prediction intervals were generally highly accurate while lacking precision. Since the intervals are so large, they capture most of the data but lack specificity in prediction.

The results indicate that prediction accuracy is limited to approximately 10% of the full dataset duration. When the models predicted further than approximately 10% of the full dataset, both models lacked specificity and defaulted to a mean function. Though the mean function often maintained accuracy in prediction, it also broadened the prediction interval, further losing precision.
Table 4.2 User 5542 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,1,1)/(1,0,1)</td>
<td>0.0439/0.0599</td>
<td>970.3096 km²</td>
<td>100%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0)/(1,2,4)</td>
<td>0.0245/0.1081</td>
<td>944.0653 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,0)/(0,1,1)</td>
<td>0.0276/0.159</td>
<td>1289.32 km²</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,1,1,0,1,2,24)/(1,0,1,0,0,2,24)</td>
<td>0.0496/0.0604</td>
<td>901.8189 km²</td>
<td>100%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0,0,1,2,24)/(1,2,4,0,0,2,24)</td>
<td>0.0365/0.1082</td>
<td>1094.65 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,1,0,2,24)/(0,1,1,1,0,2,24)</td>
<td>0.0312/0.2418</td>
<td>1821.053 km²</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 4.3 User 5975 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,3)/(0,0,3)</td>
<td>0.0320/0.00895</td>
<td>925.9497 km²</td>
<td>88.2%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,0,3)/(0,0,3)</td>
<td>0.0298/0.0910</td>
<td>907.7103 km²</td>
<td>88.2%</td>
</tr>
<tr>
<td>50%:</td>
<td>(0,0,3)/(0,0,3)</td>
<td>0.0306/0.0894</td>
<td>909.0126 km²</td>
<td>94.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,3,0,0,2,24)/(0,0,3,0,0,2,24)</td>
<td>0.0310/0.1038</td>
<td>998.9533 km²</td>
<td>82.3%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,0,3,0,0,2,24)/(0,0,3,0,0,2,24)</td>
<td>0.0304/0.1016</td>
<td>978.4469 km²</td>
<td>82.3%</td>
</tr>
<tr>
<td>50%:</td>
<td>(0,0,3,0,0,2,24)/(0,0,3,0,0,2,24)</td>
<td>0.0310/0.0955</td>
<td>949.5912 km²</td>
<td>88.2%</td>
</tr>
</tbody>
</table>
Table 4.4 Mean MSE and Prediction Accuracy by Results Group

<table>
<thead>
<tr>
<th>Results Group</th>
<th>Data Input</th>
<th>Mean MSE ARIMA ($km^2$)</th>
<th>Mean Prediction Accuracy ARIMA</th>
<th>Mean MSE SARIMA ($km^2$)</th>
<th>Mean Prediction Accuracy SARIMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Sensitive</td>
<td>90%</td>
<td>854.4704</td>
<td>91.3%</td>
<td>907.2976</td>
<td>82.8%</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>825.2287</td>
<td>94.1%</td>
<td>872.5569</td>
<td>91.1%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>887.1881</td>
<td>97.0%</td>
<td>945.853</td>
<td>85.7%</td>
</tr>
<tr>
<td>Full Data</td>
<td>90%</td>
<td>947.9128</td>
<td>90%</td>
<td>1175.7525</td>
<td>92.5%</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>1009.1005</td>
<td>92.5%</td>
<td>1071.7214</td>
<td>92.5%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>1223.3176</td>
<td>85%</td>
<td>2375.3857</td>
<td>80.35%</td>
</tr>
<tr>
<td>Total</td>
<td>90%</td>
<td>916.7653</td>
<td>90.4%</td>
<td>1086.2675</td>
<td>89.3%</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>947.8099</td>
<td>93.0%</td>
<td>1005.33</td>
<td>92.1%</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>1111.2745</td>
<td>89.0%</td>
<td>1898.8748</td>
<td>82.2%</td>
</tr>
</tbody>
</table>
CHAPTER 5  
SUMMARY AND CONCLUSIONS

As proposed in this report, an individual’s regular activities typically contain pattern and seasonality. Given recorded GPS locations through an individual’s smart phone, graphical depiction of these locations provides an opportunity for modeling past behavior and predicting future movements. By applying statistical modeling techniques, trends, seasonality and random error can be estimated. Given such parameters, the individual’s future locations can be predicted. Specifically when considering an application where the individual is conducting nefarious or illegal activities, predicted locations provide opportunities for preventing or intercepting such acts.

A substantial portion of the research conducted for this report was allocated towards data processing in preparation for applying time series methods. Since the data lacked quality in duration, accuracy, and consistency, several methods were applied to accommodate for the shortcomings. I implemented methods to truncate, process and interpolate the data for future modeling.

I modeled the individuals’ given datasets using two statistical methods: ARIMA and SARIMA models. Given time inputs of varying duration, I modeled the individuals’ past behaviors. Using the defined models, I predicted future locations with an average of 82.2% to 92.1% accuracy given a 95% prediction interval.

The most limiting factor to the output quality of this work was the sub par data. Though the LDCC was conducted over a two-year period, much of the consistent location data was less than a week in duration. Undoubtedly, this restriction limited prediction potential. Given a larger dataset over a greater duration of time, future works with the methodology outlined in this report could generate more significant results. For example, with a larger time duration for modeling, seasonality and trend could be better estimated and glean more
precise results. This would lead to longer and more insightful prediction intervals, versus the current 10% of the dataset length limitation. Additionally, predictions would have smaller intervals, providing greater precision on predicted locations. Another challenge posed with the data was the apparent errors in measured locations that were identified in the data processing portion of this report. In many instances, an individual had multiple reported locations for a given time period. Additionally, frequently an individual’s recorded location changed so drastically in such short durations of time, the velocity of travel suggested rates greater than 160 kilometers per hour.

Proposed future works also include shortening the interpolation frequency from 60 minutes to any shorter duration. By doing this, an analyst may capture more specificity in the model which would then be translated into the predictions.
REFERENCES CITED


APPENDIX A
REMAINING USERS DATA PROCESSING

Data processing consisted of truncating the original data into a concentrated region of activity, finalizing the data into valid and logical points, and interpolating locations at even intervals for further analysis. As seen in Table A.1, the truncating and final processing generally resulted in datasets approximately 10% shorter than the original set. The exception to this outcome is User 5973, whose dataset was approximately 50% shorter in the Final Processed form due to multiple locations recorded at a single time period and movement velocities exceeding 160 kilometers per hour.

Table A.1 Individual User Dataset Length for Remaining Users

<table>
<thead>
<tr>
<th>User</th>
<th>Original</th>
<th>Truncated</th>
<th>Final Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 5479</td>
<td>13709 points/7 days</td>
<td>13709 points/7 days</td>
<td>13297 points/7 days</td>
</tr>
<tr>
<td>User 5578</td>
<td>16224 points/60 days</td>
<td>15987 points/6 days</td>
<td>15127 points/6 days</td>
</tr>
<tr>
<td>User 5927</td>
<td>24346 points/29 days</td>
<td>23772 points/6 days</td>
<td>22744 points/6 days</td>
</tr>
<tr>
<td>User 5973</td>
<td>37765 points/52 days</td>
<td>34232 points/9 days</td>
<td>16057 points/9 days</td>
</tr>
</tbody>
</table>
A.1 User 5479

Figure A.1 Original versus Final Processed Data for User 5479.
Figure A.2 Original vs Interpolated 60s vs Interpolated 60m User 5479.
A.2 User 5578

Figure A.3 Original versus Truncated versus Final Processed Data for User 5578.
Figure A.4 Original vs Interpolated 60s vs Interpolated 60m User 5578.
A.3 User 5927

Figure A.5 Original versus Truncated versus Final Processed Data for User 5927.
Figure A.6 Original vs Interpolated 60s vs Interpolated 60m User 5927.
A.4 User 5973

Figure A.7 Original versus Truncated versus Final Processed Data for User 5973.
Figure A.8  Original vs Interpolated 60s vs Interpolated 60m User 5973.
APPENDIX B

REMAINING USERS DATA MODELING

B.1 User 5479

Figure B.1 Additivity Model User 5479.
Figure B.2 Autocorrelation function for User 5479.
B.2 User 5578

Figure B.3 Additivity Model User 5578.
Figure B.4 Autocorrelation function for User 5578.
B.3 User 5927

Figure B.5 Additivity Model User 5927.
Figure B.6  Autocorrelation function for User 5927.
Figure B.7  Additivity Model User 5973.
Figure B.8 Autocorrelation function for User 5973.

ACF Final Processed Latitude for User 5973

ACF Final Processed Longitude for User 5973
APPENDIX C
REMAINING USERS DATA PREDICTION

C.1 User 5479

Table C.1 User 5479 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,0)/(1,0,0)</td>
<td>0.0342/0.0606</td>
<td>782.6483 km²</td>
<td>100%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0)/(1,0,0)</td>
<td>0.0425/0.0637</td>
<td>902.898 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,0)/(1,0,0)</td>
<td>0.0447/0.0651</td>
<td>938.7632 km²</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,0,0,0,2,24)/(1,0,0,0,0,2,24)</td>
<td>0.0250/0.0556</td>
<td>638.8481 km²</td>
<td>100%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0,1,0,2,24)/(1,0,0,0,0,2,24)</td>
<td>0.0315/0.0558</td>
<td>720.2793 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,0,1,1,2,24)/(1,0,0,1,1,2,24)</td>
<td>0.1738/0.4246</td>
<td>4651.387 km²</td>
<td>71.4%</td>
</tr>
</tbody>
</table>
Figure C.1 True versus Predicted Latitude and Longitude Locations, User 5479.
Figure C.2 Final True versus Predicted Points User 5479.
Table C.2 User 5578 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(0,1,1)/(2,1,0)</td>
<td>0.0544/0.0988</td>
<td>1257.414 km²</td>
<td>70%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,1,1)/(2,1,0)</td>
<td>0.0596/0.1017</td>
<td>1339.806 km²</td>
<td>70%</td>
</tr>
<tr>
<td>50%:</td>
<td>(0,1,1)/(0,1,1)</td>
<td>0.0751/0.1288</td>
<td>1690.952 km²</td>
<td>60%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(0,1,1,1,2,24)/(2,1,0,1,1,2,24)</td>
<td>0.1212/0.0986</td>
<td>2081.336 km²</td>
<td>70%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,1,1,1,1,2,24)/(2,1,0,1,1,2,24)</td>
<td>0.0882/0.0841</td>
<td>1587.478 km²</td>
<td>80%</td>
</tr>
<tr>
<td>50%:</td>
<td>(0,1,1,0,1,2,24)/(0,1,1,0,0,2,24)</td>
<td>0.0862/0.1320</td>
<td>1846.569 km²</td>
<td>60%</td>
</tr>
</tbody>
</table>
Figure C.3 True versus Predicted Latitude and Longitude Locations, User 5578.
Figure C.4 Final True versus Predicted Points User 5578.
C.3  User 5927

Table C.3  User 5927 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,0)/(1,0,1)</td>
<td>0.0206/0.0886</td>
<td>781.2793 km$^2$</td>
<td>90%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0)/(0,1,1)</td>
<td>0.0205/0.1007</td>
<td>849.6328 km$^2$</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,0)/(1,1,2)</td>
<td>0.0202/0.1223</td>
<td>974.2355 km$^2$</td>
<td>80%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(1,0,0,0,1,2,24)/(1,0,1,1,1,2,24)</td>
<td>0.0365/0.1063</td>
<td>1081.007 km$^2$</td>
<td>100%</td>
</tr>
<tr>
<td>70%:</td>
<td>(1,0,0,1,1,2,24)/(0,1,1,1,0,2,24)</td>
<td>0.0281/0.0907</td>
<td>884.4785 km$^2$</td>
<td>90%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,0,1,1,2,12)/(1,1,2,1,1,2,24)</td>
<td>0.0521/0.0907</td>
<td>1182.534 km$^2$</td>
<td>90%</td>
</tr>
</tbody>
</table>
Figure C.5 True versus Predicted Latitude and Longitude Locations, User 5927.
Figure C.6 Final True versus Predicted Points User 5927.
### Table C.4 User 5973 Data Models by Input Percentage, MSE (Lat/Long), Total MSE, and Resulting Prediction Accuracy

<table>
<thead>
<tr>
<th>Data Input</th>
<th>ARIMA Model (Lat/Long)</th>
<th>ARIMA MSE (Lat/Long)</th>
<th>ARIMA Total MSE</th>
<th>ARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(0,1,1)/(2,1,4)</td>
<td>0.0296/0.0703</td>
<td>782.9912 km²</td>
<td>94.4%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,1,1)/(0,1,1)</td>
<td>0.0294/0.0639</td>
<td>742.7472 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,1)/(0,1,1)</td>
<td>0.0364/0.0701</td>
<td>865.3637 km²</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Input</th>
<th>SARIMA Model (Lat/Long)</th>
<th>SARIMA MSE (Lat/Long)</th>
<th>SARIMA Total MSE</th>
<th>SARIMA Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%:</td>
<td>(0,1,1,0,0,2,24)/(2,1,4,0,0,2,24)</td>
<td>0.0299/0.0751</td>
<td>815.642 km²</td>
<td>83.3%</td>
</tr>
<tr>
<td>70%:</td>
<td>(0,1,1,0,0,2,24)/(0,1,1,0,0,2,24)</td>
<td>0.0301/0.0664</td>
<td>766.666 km²</td>
<td>100%</td>
</tr>
<tr>
<td>50%:</td>
<td>(1,0,1,0,0,2,24)/(0,1,1,0,0,2,24)</td>
<td>0.0423/0.0709</td>
<td>942.1148 km²</td>
<td>83.3%</td>
</tr>
</tbody>
</table>
Figure C.7 True versus Predicted Latitude and Longitude Locations, User 5973.
Figure C.8 Final True versus Predicted Points User 5973.