8-2018

Analysis of Time-Series Data Embedded in an Experiment Design

Yinggu Bao
Clemson University, ybao@g.clemson.edu

Analysis of Time-Series Data embedded in an Experiment Design

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Statistics

---

by
Yinggu Bao
August 2018

Accepted by:
Dr. William Bridges, Committee Chair
Dr. Brook Russell, Co-Chair
Dr. Colin Gallagher
Dr. Patrick Gerard

Abstract

In this dissertation, our objective is to determine the impact of air temperature and rainfall on soil temperature with proximity to sidewalk. First, we concerned the many missing values in our data set. Some traditional and commonly used methods of imputation are introduced and then applied to the data set. Then, we concerned the relationships among the soil temperature, air temperature, and rainfall data. Since these data are time series, Granger Causality is used to estimate the relationships. Lastly, we focused on the analysis of the actual study design. The four distances are considered as treatments and the three locations were considered blocks resulting in a Randomized Complete Block Design (RCBD). Results from the time series analyses were incorporated into the study design to determine the effect of sidewalk on soil temperature.

Dedication

I dedicate this work to my loving parents, who always believe in me and support me in every step of the way.

Acknowledgments

First and foremost, I am very grateful to my advisors, Dr. William Bridges and Dr. Brook Russell for their patience and support in overcoming numerous obstacles I have been facing through my research. I could not have completed the work without their inspiration and advice.

I would like to express my gratitude to my committee, Dr. Colin Gallagher and Dr. Patrick Gerard for their encouragement and insightful comments. I would like to thank my fellow doctoral student Tianhui Wei for many discussions we had about details of research. In addition, I would like to thank all the staff in the Department of Mathematics Sciences.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Urbanization refers to the population shift from rural to urban areas, the gradual increase in the proportion of people living in urban areas, and the ways in which each society adapts to the change (Satterthwaite, McGranahan & Tacoli, 2010). Urbanization has become a powerful environmental force as the world's urban population continues to increase. One of the examples of the urbanization process is the creation of sidewalks. These sidewalks, commonly made from materials such as wood, brick, stone, or concrete provide ease of access. However, these sidewalks have caused many plants species to underperform in several aspects. For example, studies have shown that fruit plants close to concrete sidewalks tend to have reduced yield (Shi, Shao, Liu & Wang, 2009). In addition, sidewalks also cause plants to have reduced life spans. The cause of this reduced yield and life span is often assumed to be heat produced from the sidewalk resulting in a warmer soil environment compared to the environment in the surrounding area. A warmer soil environment is an issue since soil temperature affects the rate of virtually all biochemical processes in plants and is a critical determinate of plant growth. Thus, many agricultural research projects have studied the impact of urbanization (and specifically sidewalks) on the temperature profiles of plants.

Concrete is the most common material used in sidewalks in the United States and Canada. This material can be particularly harmful to plants due to temperature and also moisture impacts. This motivated us to take a closer look at a long-term concrete sidewalk and soil temperature data

set and develop a method to carefully extract all the useful information available on the impact of the sidewalk on temperature. The findings from this research can help agricultural researchers to further optimize landscape management to maximize plant yield or production.

1.2 Data Description

Soil environmental monitoring sensors are valuable tools used to record soil temperature in landscape, nursery and agricultural management. Previous research has been conducted using soil environmental monitoring sensors to assist in irrigation management in greenhouse production, sports fields, golf courses, and residential landscapes (Ersavas, 2014). A study was conducted using GS3 soil moisture sensors in the Sustainable Demonstration Garden (Figure 1.1) at Clemson University to determine the impact of sidewalks on soil temperature (Drennan et al., 2014). In this study, three separate locations were chosen along the sidewalk in the Sustainable Demonstration Garden at Clemson. Within each location, GS3 soil sensors were installed at 15 cm, 30 cm, 45 cm and 60 cm increments from the sidewalk, at a depth of 10 cm into the soil on August 9, 2013 (Figure 1.2). In this dissertation, these three locations are considered as replications (denoted as "rep") or blocks. These three reps were at a distance greater than 45cm from existing plants, were in what appeared to be undisturbed sites, and were adjacent to a sidewalk. Therefore the only influence on the temperatures should be the sidewalk, not other factors. Mulch was moved to the side and a narrow 10 x 80 cm trench was dug. The soil was then replaced, uniformly compacted similar to the surrounding soil and remulched. Measurements from the sensors were logged every 30 minutes using EM50G data loggers and uploaded to a remote server managed by Decagon Devices from August 2013 to August 2015. Precipitation and air temperature data recorded at every 30 minutes interval from the same time period were downloaded from the Clemson University Entomology Weather Station.

1.3 Overall Objective and Dissertation Organization

The overall objective of the study was to determine if proximity to the sidewalk influences soil temperature. The specific hypothesis was that the sensors near the sidewalk would have higher soil temperature profiles than sensors placed further from the sidewalk. In addition, we would like

Figure 1.1: Sustainable Demonstration Garden

Figure 1.2: Sensor Installation

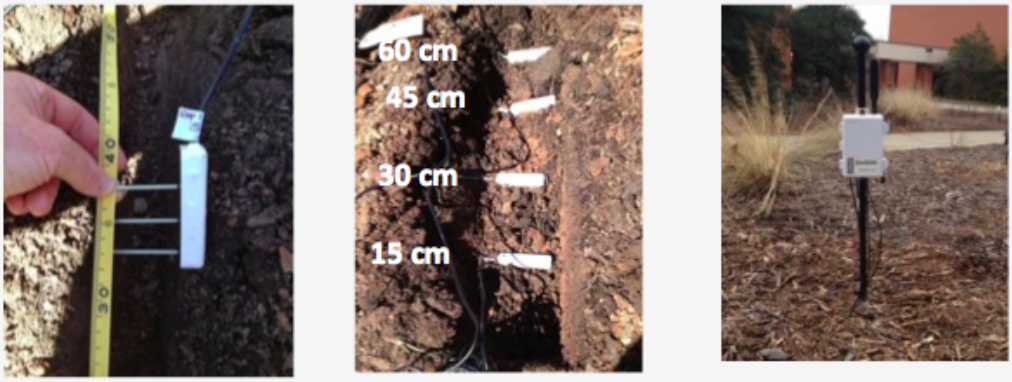


Fig. 1 A close up of the GS3 soil sensor

Fig. 2 Sensors were installed at 15 cm increments from the sidewalk

Fig. 3 Data logged every 30 min using a data logger

to characterize the impact of air temperatures and rainfall on the influence of sidewalks on soil temperature. The findings from this study will provide important information to growers so that they can make good management decisions involving sidewalks and plants, specifically decisions such as irrigation practices, plant selection and/or plant placement.

The remainder of this dissertation is organized as follows. Chapter 2 concerns the many missing values in our data set. Some traditional and commonly used methods of imputation are introduced and then applied to the data set. Chapter 3 concerns the relationships among the soil temperature, air temperature, and rainfall data. Since these data are time series, Granger Causality is used to estimate the relationships. Chapter 4 focuses on the analysis of the actual study design. The four distances are considered as treatments and the three locations were considered blocks resulting in a Randomized Complete Block Design (RCBD). Results from the time series analyses were incorporated into the study design to determine the effect of sidewalk on soil temperature. The results of the data analyses and a summary discussion are also provided in Chapter 4.

Chapter 2

Imputation of Missing Values in Time-Series Data Sets

2.1 Introduction

Missing values are a problem that frequently occurs in data collection processes. In our soil temperature time series dataset, there are many missing values. Various reasons could result in missing values including: 1) values may not be measured, 2) values may be measured but get lost, or 3) values may be measured but are considered unusable. In the soil temperature data set the reason for the missing values was typically that the sensor had a malfunction at certain points in time. Missing values can lead to problems, because many data processing and analysis steps often rely on complete datasets. One approach to overcoming these problems is to replace the missing values with estimated values. In statistics this process is called imputation.

Time series data sets often have missing values. Before discussing imputation techniques for estimating the missing values in time series, we will review some common definitions and concepts in time series data sets. A time series data set is a type of data set in which observations of a response variable of interest (often denoted Y) are observed on a regular interval of successive times $(t_1, t_2, t_3, \cdots, t_n)$. A univariate time series is a sequence of single observations at successive points $t_1, t_2, t_3, \cdots, t_n$ in time. Although a univariate time series is usually considered as one column of observations, time is in fact an implicit variable. In this study, we only considered univariate time

series with equispaced time intervals meaning that time increments between successive data points are equal, $|t_1 - t_2| = |t_2 - t_3| = ... = |t_{n-1} - t_n|$. Two common approaches to describe and examine time series data set are autocorrelation analysis and separation into trend, seasonal and irregular components (Moritz, Bartz-Beielstein, Zaefferer & Stork, 2015).

2.1.1 Autocorrelation

The analysis of a time series data set requires different considerations than those generally encountered in more traditional data analysis procedures. The distinguishing aspect of the structure of time series data is the non-independence of Y observations. Most statistical models are based on the assumptions that the Y observations are independent, or uncorrelated, with other observations. However, this basic assumption is seldom satisfied for observations that are collected across time (unless they are measured on different experimental units). Instead, Y observations are likely to be related to other Y observations collected in close temporal proximity (and hopefully relatively independent from more distant observations). For example, in our soil temperature data, soil temperature for yesterday is probably related with today's soil temperature. But soil temperature collected a month ago may not be related with today's soil temperature. So autocorrelation, also called serial correlation, is used to measure the correlation of Y observations within a time series.

The autocorrelation of a time series process is defined as the correlation between all pairs of observations that are separated by a fixed number of points in the time series. It is a representation of the degree of similarity between the time series and a lagged version of itself. The calculation process is similar to getting correlations between two different data series, except that one time series is used twice, once in its original form and once in a lagged version. Suppose that some soil temperatures are measured on a monthly basis over 2 years time period. The estimated correlations between the soil temperature on month 1 vs. month 2, month 2 vs. month 3,$\cdots$, through month 23 vs. month 24 are the first-order autocorrelations. This first-order autocorrelation coefficient is an indication of how well the overall monthly soil temperature can be predicted on the basis of the soil temperature on the previous month. Similarly, the h-order autocorrelation coefficient (or autocorrelation at a lag of h) can be computed by correlating observations on month t vs. month t+h.

The h-order autocorrelation, that is the correlation between the two random variables Y_t

and Y_{t+h} in a time series, is defined as:

$$\rho(h) = \frac{Cov(Y_t, Y_{t+h})}{Var(Y_t)} = Corr(Y_t, Y_{t+h}). \tag{2.1}$$

where Y_t is the observation at time t, Y_{t+h} represents the observation at time t+1. The numerator in the equation of autocorrelation $Cov(Y_t, Y_{t+h})$ is the covariance between Y_t and Y_{t+h} of a time series, which is called autocovariance at lag h.

An estimator or sample autocorrelation is used to estimate the population autocorrelation and is defined as

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{2.2}$$

where n is the number of observations and $\bar{y}$ is the mean of all the observations.

The basic idea of using autocorrelation in imputation of a time series is that future observations usually depend on past observations. High autocorrelation values indicate the future is strongly correlated to the past. Thus autocorrelation can be used as an indicator for imputation reliability.

The range of the sample autocorrelation is from -1 to +1. A value of +1 means that there is a perfect positive association, a value of -1 means that there is a perfect negative association and zero means there is no association.

2.1.2 Decomposition

Time series data usually show a variety of patterns and, for analyses, it can be useful to isolate these patterns in to separate components of the series. Time series decomposition seeks to split the time series into single components each representing a certain characteristic or pattern. The original time series can later on be reconstructed by additions or multiplications of these components. There are typically three components of interest:

- *Trend Component:* expresses the long-term progression of the time series. This means there is a long-term increase or decrease in the mean level of the Y observations. The trend does not necessarily have to be linear.

- *Seasonal Component:* is a common pattern repeating, for example every month, quarter of the year, or day of the week. Seasonality is always of a fixed and known period, i.e. season length is fixed. And there are two types of seasonalities: additive and multiplicative. For example, the average soil temperature in June may rise by 6 °C in comparison to May. Thus, the amount of 6 °C over the average soil temperature in May was added to predict for every June to account for this seasonal fluctuation. In this case, the seasonality is additive. Alternatively, during the month of June the soil temperature may increase by 20%, that is, increase by a factor of 1.2. In this case the temperature increase by a certain factor, and the seasonal component is thus multiplicative in nature (Moritz et al., 2015).

- *Irregular Component:* describes irregular influences. These are the residuals, after the other components have been removed from the time series. The irregular component may be, but is not necessarily completely random.

Considering trend and seasonal influences is very important for the analyses of time series data. Getting trend and seasonal effects modeled correctly can improve imputation results considerably. Thus decomposition is a popular technique.

2.1.3 Missingness Mechanism

For time series data (or really any type of data), missing observations can occur due to several reasons that are known as the "missingness mechanism". Missingness mechanism can be divided into three categories: Missing Completely At Random ($MCAR$), Missing At Random (MAR) and Not Missing At Random ($NMAR$).

To better understand and describe the following mechanisms, some notations about missing values are introduced based on (Rubin, 1976). Let $Y_i, i = 1, \cdots, n$ denote a complete series of data which contains both the observed Y_{obs} and the missing values Y_{mis}. $M = (M_1, ..., M_n)$ is the missing data indicator variable which denotes whether the value of a variable is observed or missing (i.e. $M_i = 0$ if value Y_i is observed and $M_i = 1$ if the value is missing). The pattern of missing data is defined by the missing data indicator M. There is an underlying probability distribution of the missing data indicator when the missing data are presented as a variable. In real situation, it is impossible to know the exact distribution of M. However, the relationship between the data and missing value indicator M is used to classify the missing data mechanisms.

Missing completely at random (MCAR)

In MCAR, there is no systematic mechanism for the way the data are missing. A variable is missing completely at random if there are no dependencies in the missingness probability. There are two requirements for this to be true. First, the probability that observations from one variable are missing is independent from the values of all other variables. Second, the probability for an observation being missing is also independent of the variable itself. Since no other variables exist for univariate time series (except time as implicit variable), requirement one can be simplified to: the probability that certain observation being missing is independent of time. So the probability of an observation being missing is independent of the point in time it has been observed in MCAR (Moritz et al., 2015). In our soil temperature data sets, the soil temperature data were monitored by the soil sensors from the location and sent to the data logger. Due to unknown reasons, sometimes the sensor did not work or the transmission failed. The probability for one soil temperature being missing is independent from other soil temperatures. Also, there is no relationship between the occurrence of missing soil temperature and the value of this soil temperature. So the missingness mechanism for our soil temperature data sets can be treated as MCAR.

$$P(M|Y_{obs}, Y_{mis}) = P(M)$$

Missing at random (MAR)

Like in MCAR, in MAR the probability for an observation being missing is also independent of the value of the observation itself. But it is dependent on other variables. Since in the case of univariate time series, time is considered as an implicit variable, it can be said, that in MAR the probability for an observation being missing is dependent of the point in time of this observation in the series. For example, observations sensor data are more likely to be missing on weekends since no one is monitoring the system on weekends.

$$P(M|Y_{obs}, Y_{mis}) = P(M|Y_{obs})$$

Not missing at random (NMAR)

NMAR observations are not missing in a random manner. The missing observations are neither MCAR nor MAR. That means, the probability for a observation being missing depends on the value of the observation itself. Furthermore the probability can (but may not necessarily) be dependent on other variables (point of time in the series). For example, temperature sensor gives no values for temperatures over 30 °C.

$$P(M|Y_{obs}, Y_{mis}) = P(M|Y_{obs}, Y_{mis})$$

In practice, the actual missingness mechanisms resultig in the missing observations are often unknown. Some statistical techniques have been developed to check the type of missingness mechanism. Data can be checked for the MCAR mechanism with the Little's test (Little, 1988). In (Jamshidian, Jalal, & Jansen, 2014), some additional methods of checking for the MCAR mechanism can be found. Checking for the MAR and NMAR requires manual analysis of the patterns in the data, and application of domain knowledge. With MCAR mechanism, both observed and missing values should have the same mean and variance since they follow the same population. Dixon (1988) developed a two sample t-test to compare (simulated) missing values and the observed values. In this test, the means of missing values Y_{mis} and observations Y_{obs} are tested to examine whether they are significantly different. However, the means of missing values Y_{mis} cannot be computed for real data since no actual values are known for missing data. Therefore this method is appliable to the missing values created by a simulation, some values are artificially removed to simulate missing values. The details of this simulation are described in Section 2.3.2. In this test, when the null hypothesis is true, the test statistic follows a Student's t distribution since the variance is not known in most cases. The variances are often assumed to be not equal. The missingness mechansim is assumed to be MCAR when the p-value is greater than the significant level, and assumed to be MAR or NMAR if p-value is less than the significant level. The t-statistic is given by:

$$T = \frac{\bar{Y}_{obs} - \bar{Y}_{mis}}{\sqrt{S_1^2/n_1 + S_2^2/n_1}} \quad \sim \quad t_\nu, \tag{2.3}$$

where $\bar{Y}_{obs}$, S_1^2 and n_1 are the sample mean, sample variance and sample size for the observed data. $\bar{Y}_{mis}$, S_2^2 and n_2 are the sample mean, sample variance and sample size for the missing data. The degrees of freedom ν associated with this variance estimate is given by the Welch-Satterthwaite equation (Welch, 1974):

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Enders (2010) pointed out the disadvantages of the t-test approach. For very small group sizes, statistical power is decreased. MAR and MNAR mechanisms can produce missing data with equal means of observations and missing values.

The majority of missing data imputation methods require the missingness mechanism to be MAR or MCAR because the specific mechanism is said to be *ignorable* for them (Rubin, 1976). MAR also has the advantage of using correlations with other variables in the data set for imputation since there are some relationship between missing values and observations from other variables. NMAR is called non-ignorable, because in order to do the imputation a special model for why data are missing has to be developed. For time series data the imputation algorithms do not need to rely solely on correlations with other variables for missing value estimation, they can also use time series characteristics for the estimation. This makes estimating missing values for time series somewhat easier. Imputation for time series with MAR and MCAR mechanisms are nearly the same (Moritz et al., 2015).

In most studies there are multiple mechanisms causing missing data. For example, sensor recording failure will be treated as a missing completely at random since it has no dependencies to the observations that are missing, but a sensor that stops working and creates missing observations due to battery life will be treated as not missing at random. Several methods have been introduced to estimate missing values according to specific missing mechanisms or even a general solution for any missing values mechanism. Next we will discuss several methods for handling missing values.

2.1.4 Conventional method

Many methods have been suggested to deal with missing values. Several of them are really simple methods. Some of the methods use statistical principles as their base.

- Ignoring: The first way to deal with missing values is ignoring the missing value, which is the simplest way to deal with missing values. This method can be used under any missingness mechanism. However, if the percentage of missing values is large, this approach can severely impact the result of the analysis (Little & Rubin, 1989). Analysis results for data set containing missing values can significantly differ from those without missing values.

- Deletion: This method is to simply delete the entire observation (i.e., case or unit) with missing value and can be used with MCAR missingness mechanism. The disadvantage of this method is reduced power due to reduced sample size. There can also be bias in the results if the mechanism is not MCAR. Enders (2010) states that the disadvantages of deletion far outweigh any advantage gain by creating a complete data set.

- Mean Imputation: While ignoring and deletion do not usually result in an improved analysis, the Mean Imputation method often result in an improved solution. With mean imputation method, each missing value Y_{mis} is replaced by the overall mean. One disadvantage to this method is the possible bias caused by many observations having the same values.

In addition to the simple methods above, many other more advanced methods of imputation have been developed. Examples of popular techniques include Multiple Imputation (Rubin, 1987), Nearest Neighbor (Vacek & Ashikaga, 1980) and Hot Deck (Ford, 1983) methods . In the literature of imputation, time series data sets are a special challenge. Most of the sufficiently developed conventional methods rely on correlations among variables to estimate missing values. In time series, the correlations with previous values in the time series are as important, if not more important, as that correlation with additional variables. Therefore, effective algorithms need to make use of the time series characteristics. This has resulted in the development of imputation algorithms especially tailored for time series. In the case of univariate time series, imputation methods need to exploit time series characteristics, in order to estimate the values of the missing data (Moritz & Bartz-Beielstein, 2015). Since our soil temperature time series data is seasonal, we need to apply methods that perform well for seasonal time series data.

The remainder of this Chapter is organized as follows. Section 2.2 introduces some advanced imputation methods. Section 2.3 compares 4 imputation methods for simulated missing values. We applied seasonal Kalman filter method to our soil temperature data set and the results of the imputation are in Section 2.4. Section 2.5 concludes with a summary discussion.

2.2 Advanced Imputation method

This section describes some popular and commonly used imputation methods for univariate time series data.

- Replaced by the last observed value: This method is a special case of mean imputation, where each missing value Y_{mis} is replaced with the mean defined as the most recent observaion Y_{obs} prior to it. In other words, for each observation, we replace the missing value with the last observed value of that variable (Zeileis & Grothendieck, 2005). This is the simplest imputation method that takes advantage of the potential relationship between concurrent observations in time series. However, this method is problematic when there is a large time gap between a missing value and the previous non-missing value, or when the time series has seasonal (or other) patterns.

- Seasonal Kalman filter: This method uses a seasonal Kalman filter to estimate missing values Y_{mis}. The time series has to have a seasonal pattern (Harvey, 1990). The Kalman filter basically attempts to find estimates of the missing values at time t by computing the conditional mean and variance of the distribution for unobserved values conditional on observations up to time t (Durbin & Koopman ,2012).

- Interpolation: This method uses linear interpolation for non-seasonal series and a periodic decomposition with seasonal series to replace missing values (Hyndman, 2014). The seasonal component is removed from the time series in the first step, on the remaining component (trend and irregular) a linear interpolation is done to estimate the values. Afterwards the seasonal component is added again. This method is supposed to be a good fit where a clear and strong seasonality can be expected.

2.3 Imputation Implementation

In order to find the most suitable method to complete the imputation of the soil temperature data set, we considered 4 methods: the mean imputation method introduced in section 2.1.4 and the 3 imputation methods introduced in section 2.2.1. We wanted to evaluate the performance of these 4 methods in term of RMSE and MAPE (introduced in 2.3.1). Unfortunately these measures cannot

be computed for real data since no actual values are known for missing data, and we cannot calualate the difference required for these measures. Therefore we decided to perform a simulation to compare the four methods. The simulation basically involved creating a complete time series and artificially removing some values to simulate missing value. Then the missing values are estimated (using the 4 imputation methods) and the differences in the actual and imputed values can be calculated.

2.3.1 Imputation Evaluation

In this dissertation, the square root of the mean square error (RMSE) and the Mean Absolute Percentage Error (MAPE) were used to determine if the results of imputation were useful. The RMSE is a very common measure of difference between imputed and actual values. The reason for including the MAPE is that it can be useful for datasets with a strong trend (Swanson, Tayman & Bryan, 2011). For example, suppose a time series starts with very low values and ends up with very high values (ie., a strong trend exists). The missing observations near the end of the time series would have a large impact on the RMSE, while the missing observations earlier in the time series would only have a small impact. In such cases, an error measure based on the difference between imputed value and real value expressed as a percent can be more useful. Depending on specific application, either the RMSE or MAPE can best represent the quality of the imputation methods. Hence, we recorded both metrics.

RMSE The Square Root of the Mean Square Error (RMSE) between the imputed missing value $\hat{Y}_{mis}$ and the respective true missing value time series Y_{mis} , i.e.,

$$RMSE(\hat{Y}_{mis}, Y_{mis}) = \sqrt{\frac{\sum_{t=1}^{N}(Y_{mis}(t) - \hat{Y}_{mis}(t))^2}{N}}, \tag{2.4}$$

where N is the number of missing values.

MAPE The Mean Absolute Percentage Error (MAPE) between the imputed value $\hat{y}$ and the respective true value time series y, i.e.,

$$MAPE(\hat{Y}_{mis}, Y_{mis}) = \frac{\sum_{t=1}^{N} \frac{|Y_{mis}(t) - \hat{Y}_{mis}(t)|}{|Y_{mis}(t)|}}{N}, \tag{2.5}$$

2.3.2 Simulation Details

The characteristics of the algorithm created for the simulations are shown below:

- Complete Data: a complete series of soil temperature observations chosen from our original soil temperature data
- Missing Data Mechanism: MCAR, similar missing patterns of the actual soil temperature
- Missing Data Distribution: Bernoulli
- Percentage of Missing Data: adjustable

The percentage of missing values represents the parameter p of the Bernoulli distribution. The probability mass function of the Bernoulli distribution is $f(x) = p^x(1-p)^{1-x}$, $x \in (0, 1)$. We created missing values at 4 different percentages (10%, 30%, 50%, 70%). For the same percentage of missing values, the pattern of missing values can be slightly different. So we ran 30 different random seeds to randomize the results. Overall, the simulation was performed for 30 random seeds, 4 different percentages of missing values and 4 imputation methods. That equate to 480 runs for the data set.

We decided to use one real complete soil temperature data set with 100 observations chosen from our original soil temperature data as the complete data set. We took a three steps approach to compare results. In the first step, we randomly deleted some values from the complete data set (MCAR) and obtained an incomplete data set with 4 different percentages of missing values (10%, 30%, 50%, 70%). In the second step, we applied the 4 imputation methods to the incomplete time series data set. For the last step, we compared the difference between the imputed values and the actual values using RMSE and MAPE. We used the statistical software R (package "ImputeTS") to complete the imputaion with 4 different imputation methods. Figure 2.1-2.4 provide the imputation results for one random seed with different missing value percentages. Table 2.1 and Table 2.2 provide results for that one random seed.

Figure 2.5 and Figure 2.6 show the RMSE and MAPE results for 30 simulation runs and indicate comparison of different imputation methods for the real complete soil temperature data set. Each point in the figure is equivalent to one imputation result (given as MAPE or RMSE) for one variation of the time series (same complete series but with different missing values due to different random seeds). The colors in the figures mark different percentages of missing values.

In order to choose the most suitable imputation method for our actual soil temperature time series data, we then created some missing values with similar missingness patterns as the actual soil temperature data. From Figure 2.9, we can see the missingness patterns of our actual soil temperature data. Since the missingness pattern in rep2 is same as the first part of the missingness pattern in rep3. So we created another 2 missingness patterns: one is similar to the missing pattern in rep1, one is similar to that in rep3. Figure 2.7 and Figure 2.8 show the imputation results for one simulation with different missing value percentages and similar missingness patterns as the actual soil temperature.

The complete soil temperature data set we used to compare imputation methods has a clear seasonality. From Figure 2.5 , Figure 2.6, Figure 2.7, and Figure 2.8 we can find, that seasonal Kalman filter method show the best results for all missing percentages. This is probably because it can handle seasonality in the data better than the other methods. Mean imputation method shows the poorest results since the data have huge differences in the mean level. The other two imputation methods are located in the middle between this two poles. As can be seen by looking at the two tables, RMSE and MAPE lead to the same results. Depending on the missingness pattern of our soil temperature data set, the seasonal Kalman filter method was applied to complete the imputation. The results are shown in Section 2.4.

2.4 Imputation Results

In our soil temperature data, we used two year time periods, from 9/1/2013 to 8/31/2015. The data were recorded every 30 minutes. Therefore, in theory we should have 35,040 soil temperature observations (2 years x 365 days x 48 soil temperature observations per day) for the 12 combinations of locations (rep1, rep2, rep3) and distances (15cm, 30cm, 45cm, 60cm). However, we only have around 20,000 observations for distances in rep 1 (there is nearly 15,000 missing observations). Furthermore, for rep 2 and rep 3, there is about 6,000 and 10,000 missing observations, respectively. The plots of the soil temperatures over time for the 12 combinations are presented in Figure 2.9. In the plots, the missingness patterns and the seasonal components are clearly shown.

Table 2.1: RMSE for different imputation methods with different missing percentages for one simulation

RMSE				
	Missing Percentage			
	10%	30%	50%	70%
Mean	0.825	0.917	0.876	0.842
Last Observation	0.201	0.312	0.381	0.411
Kalman Filter	0.052	0.061	0.071	0.078
Interpolation	0.162	0.179	0.184	0.212

Table 2.2: MAPE for different imputation methods with different missing percentages for one simulation

MAPE				
	Missing Percentage			
	10%	30%	50%	70%
Mean	2.519	2.497	2.239	2.243
Last Observation	0.423	0.515	0.758	0.937
Kalman Filter	0.108	0.121	0.146	0.228
Interpolation	0.290	0.313	0.343	0.412

One important issue to resolve was if the imputation should be done on a monthly basis. Figure 2.10 shows the imputation results on a monthly basis. Since we have a 2 year data set, we have in total 24 months.

2.5 Discussion

From all these figures, we can know that this imputation results perform well because the imputed soil temperature for the 12 different combinations of rep and distance all have a similar seasonal pattern. In the future, we need to consider how to use the other information like the other locations, the other reps, the other time series to improve the imputation. Also, we will try to complete the imputation of the missing data in the original scale of our soil temperature time series data.

Figure 2.1: Imputed Soil Temperature with 10% as missing percentage

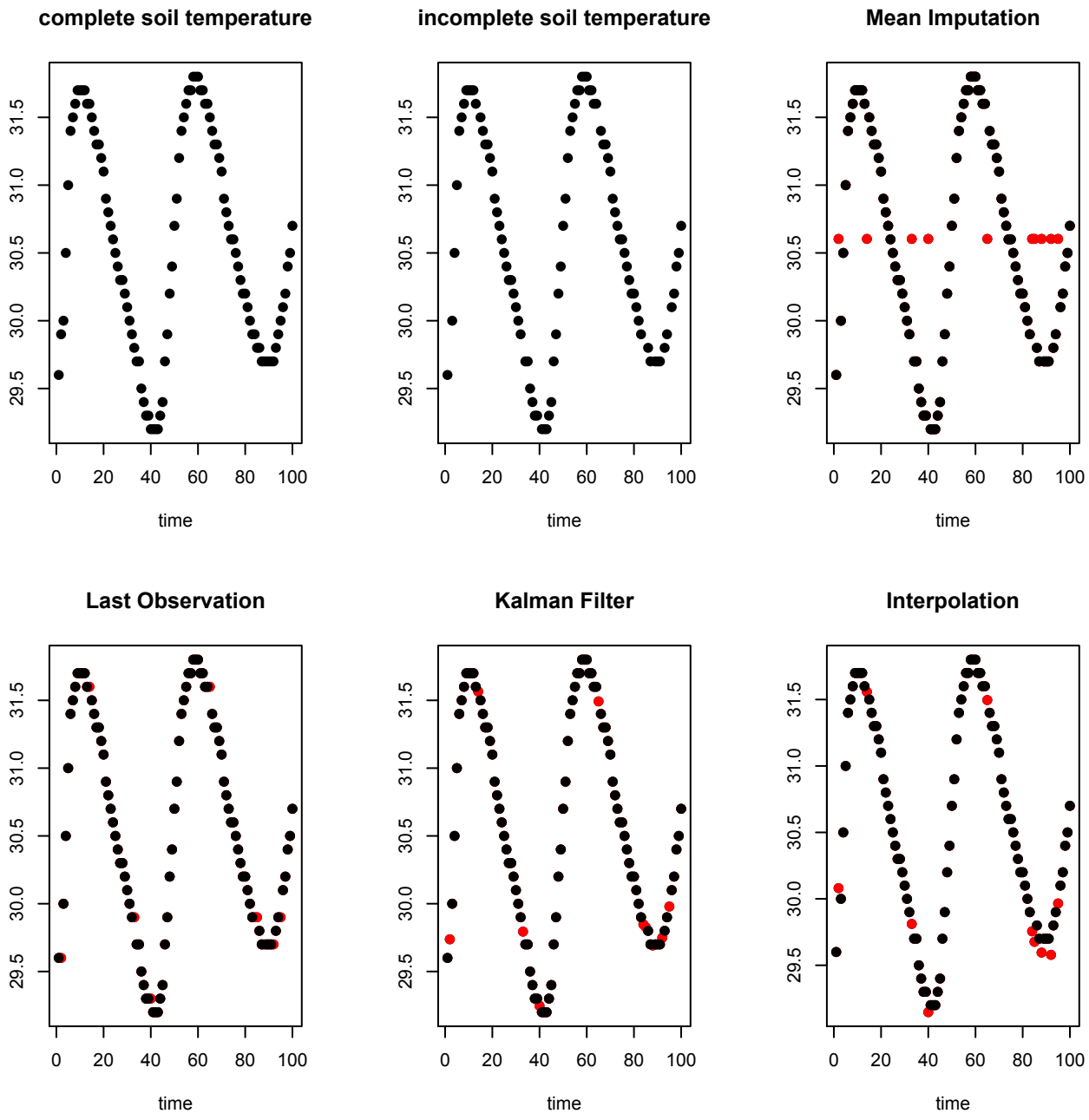


Figure 2.2: Imputed Soil Temperature with 30% as missing percentage

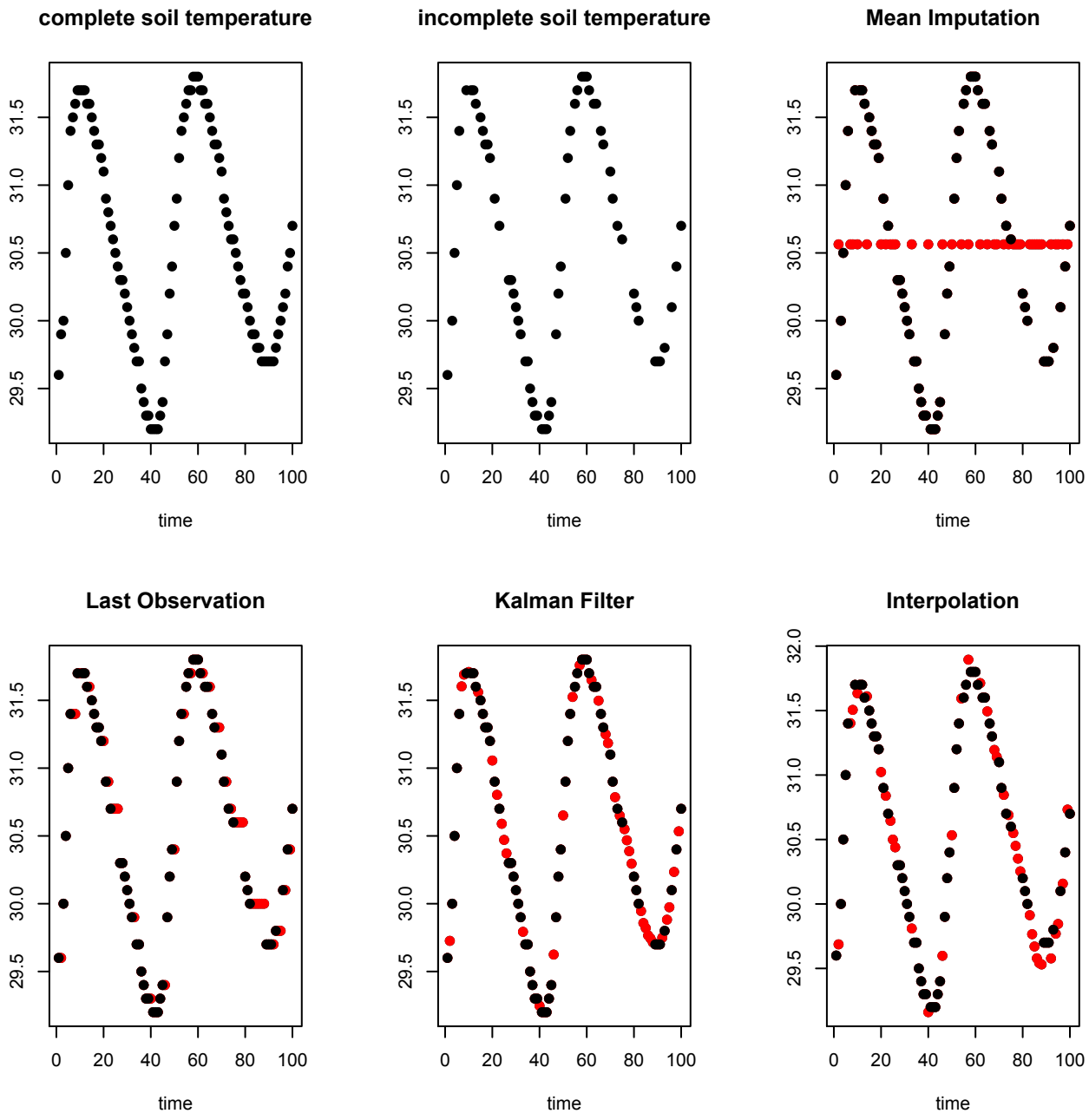


Figure 2.3: Imputed Soil Temperature with 50% as missing percentage

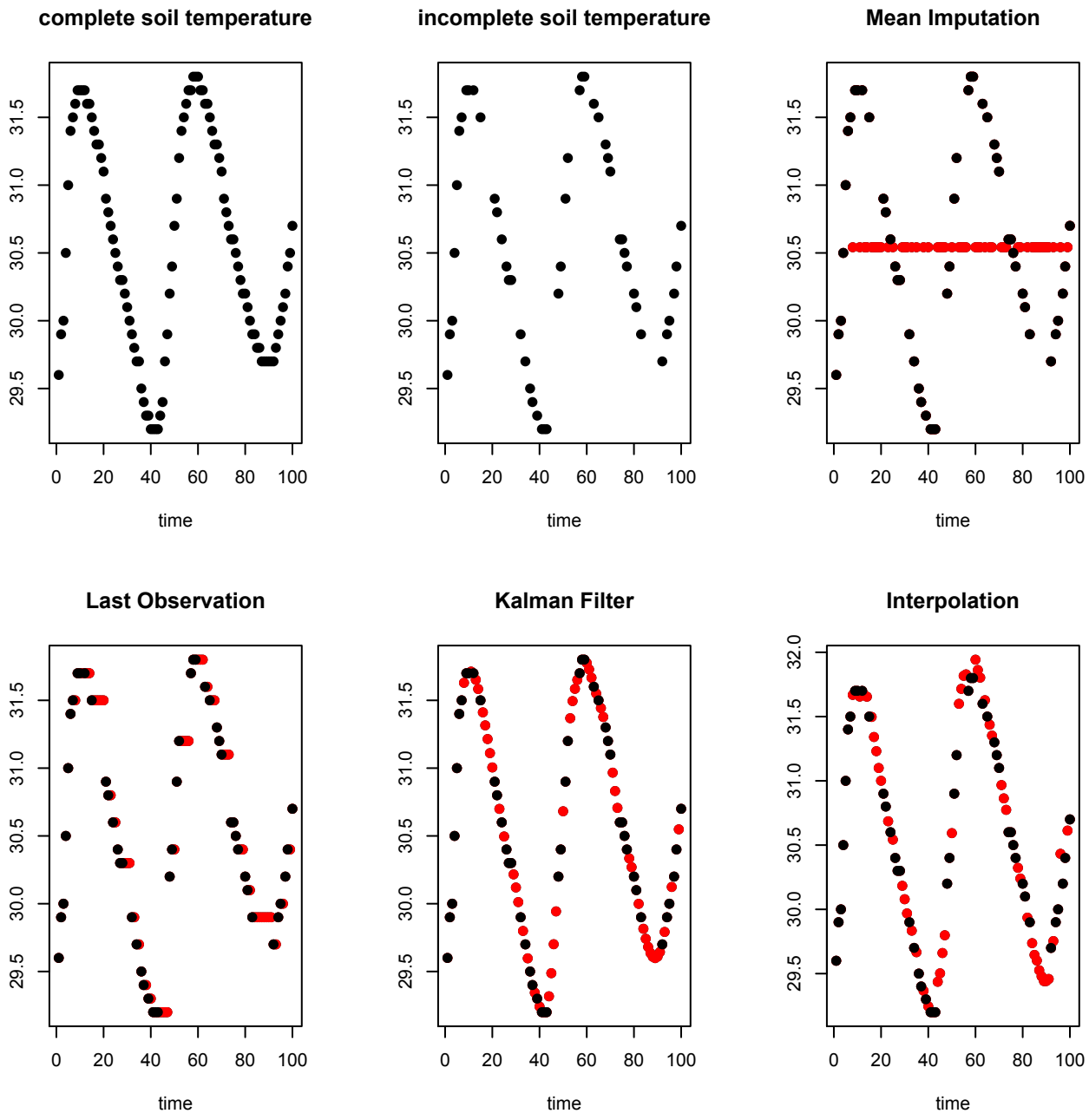


Figure 2.4: Imputed Soil Temperature with 70% as missing percentage

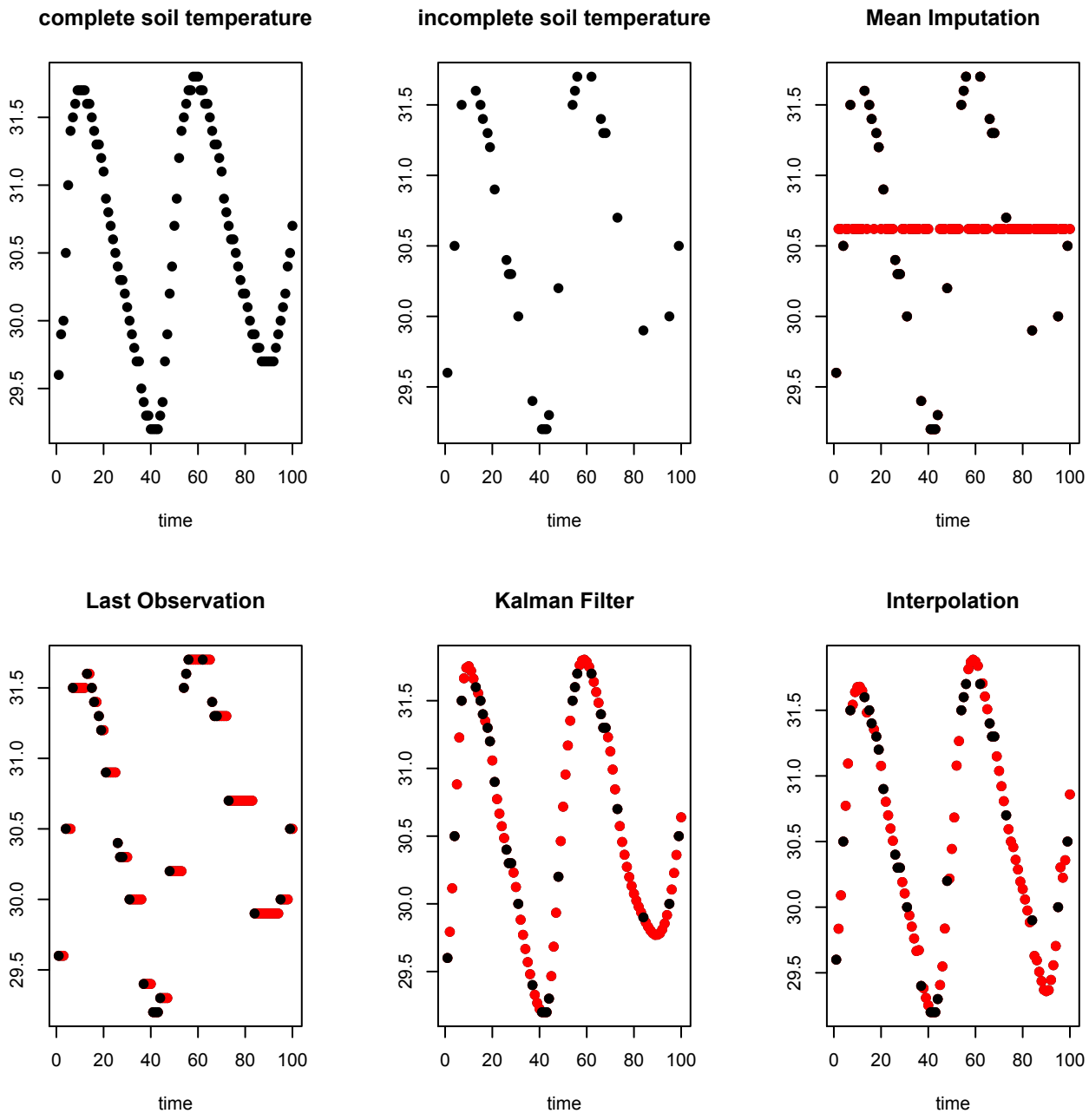


Figure 2.5: RMSE Imputation Results

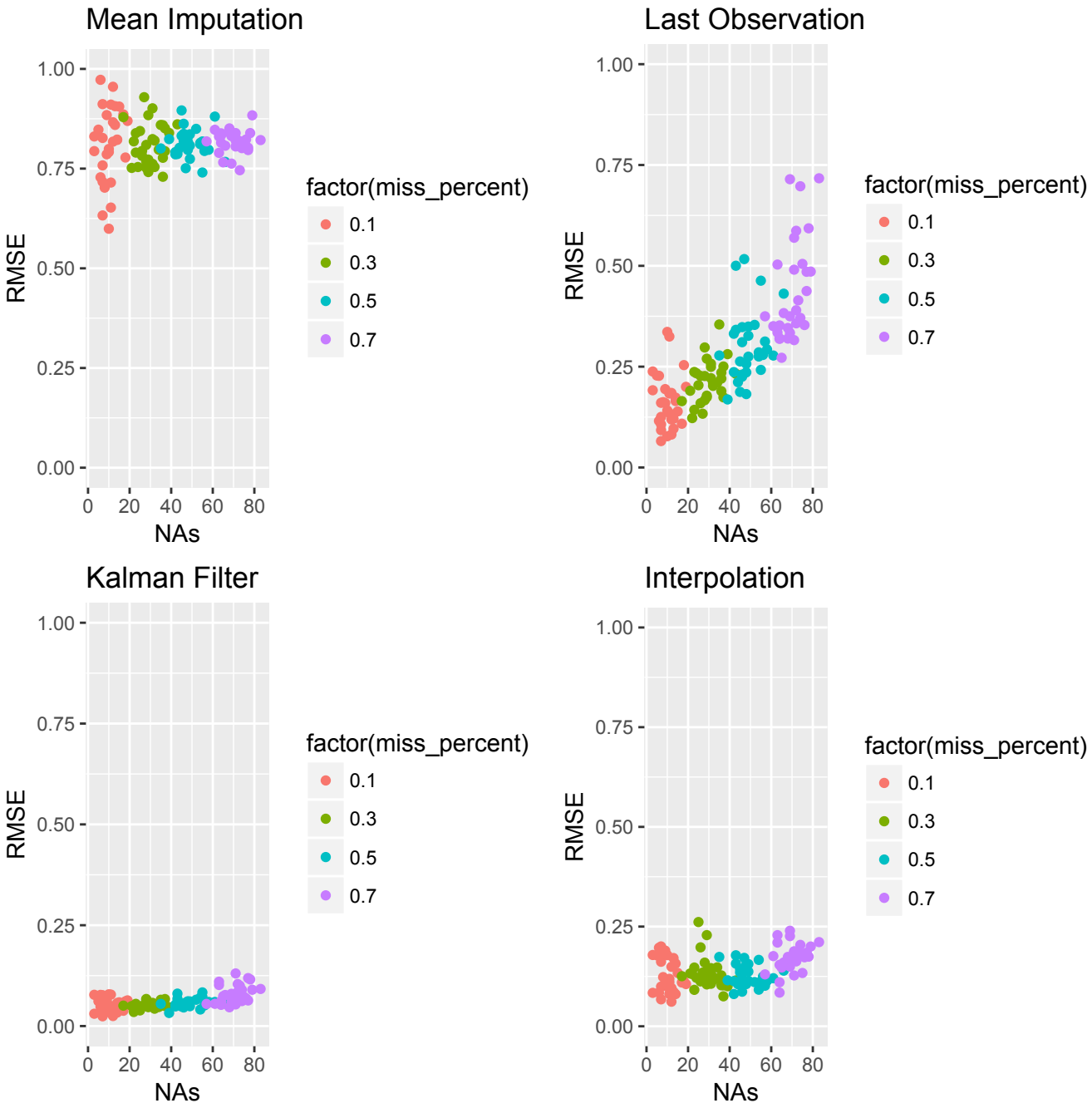


Figure 2.6: MAPE Imputation Results

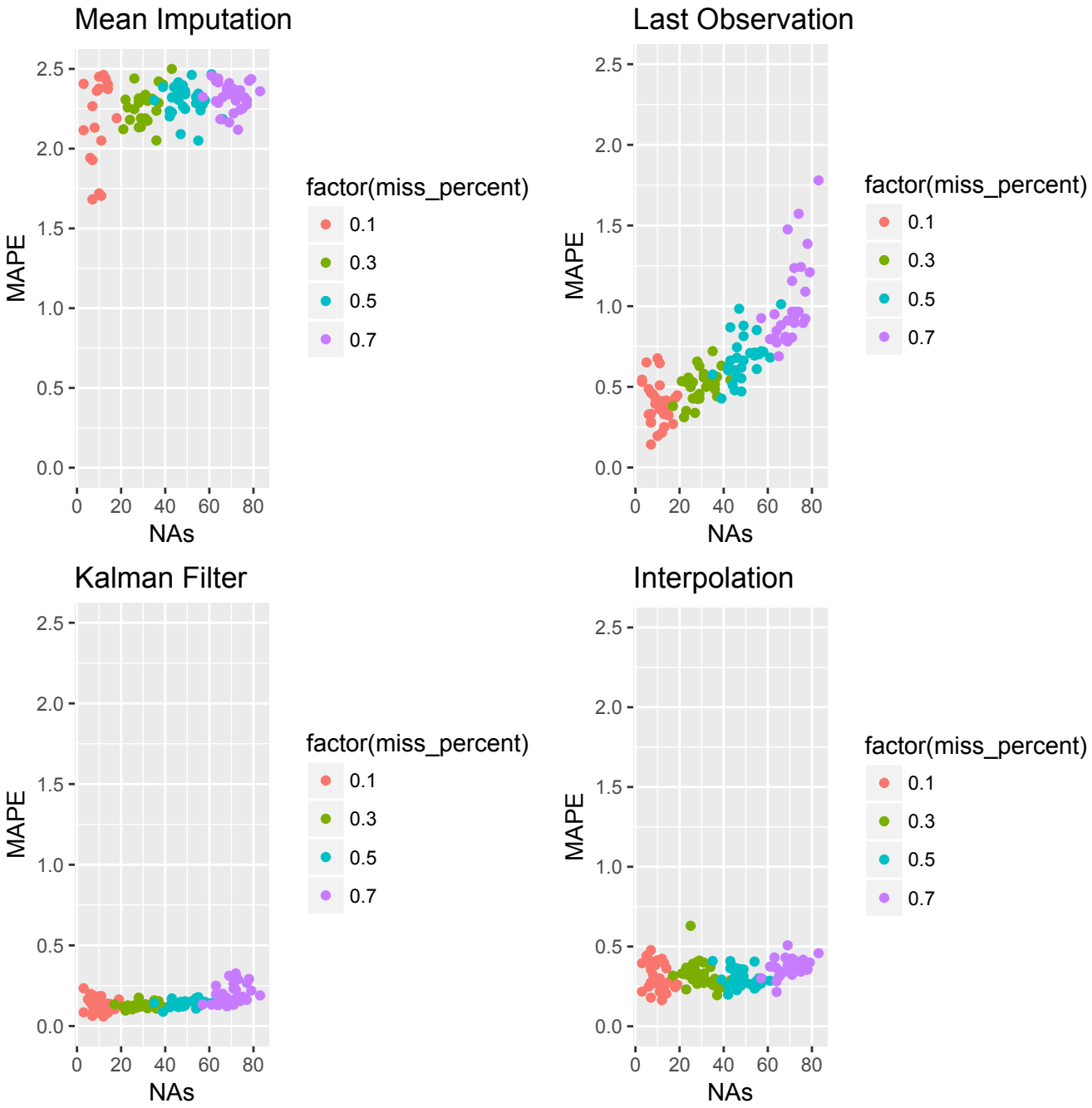


Figure 2.7: Imputed Soil Temperature with similar missingness pattern as rep 1

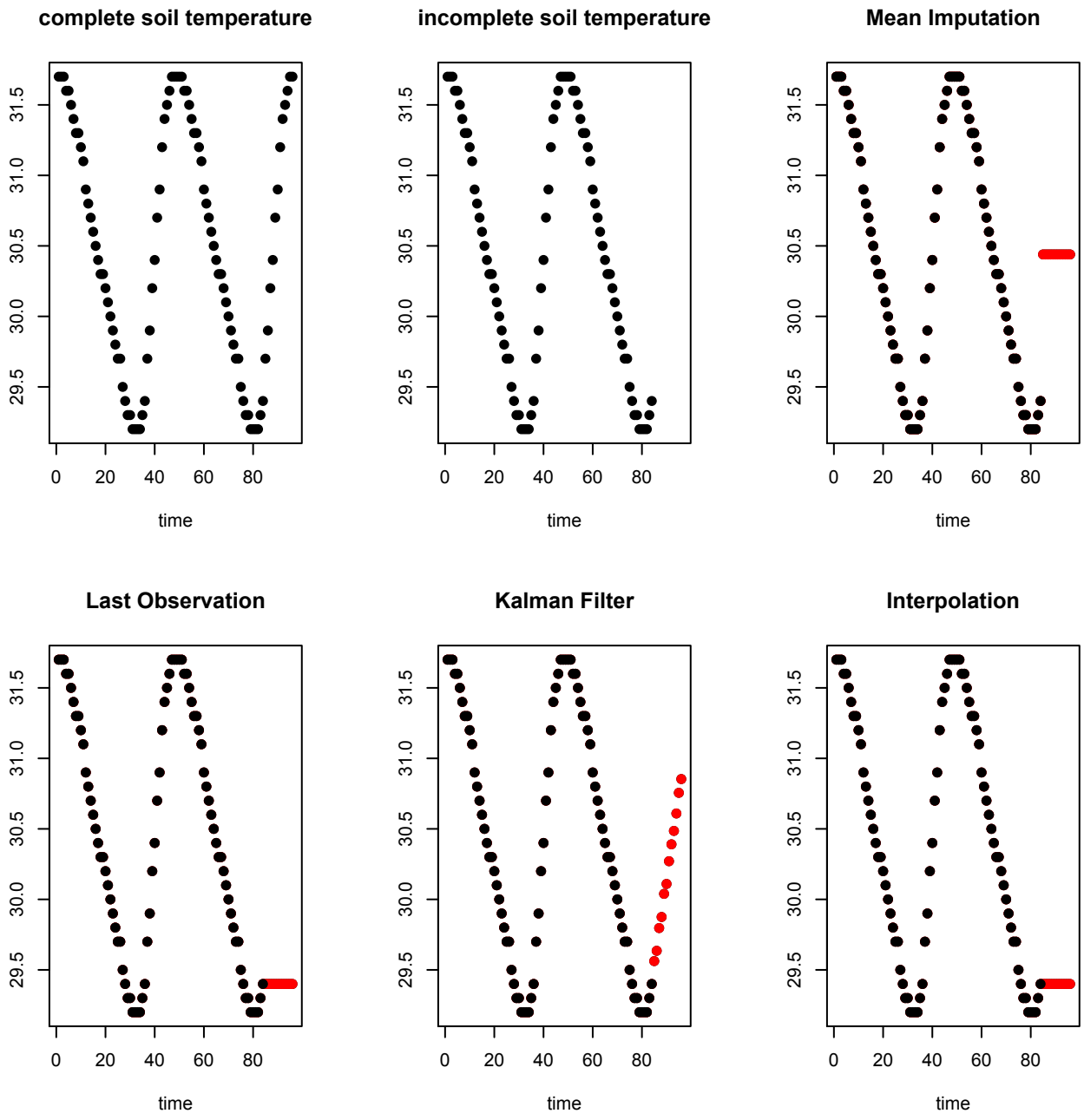


Figure 2.8: Imputed Soil Temperature with similar missingness pattern as rep 3

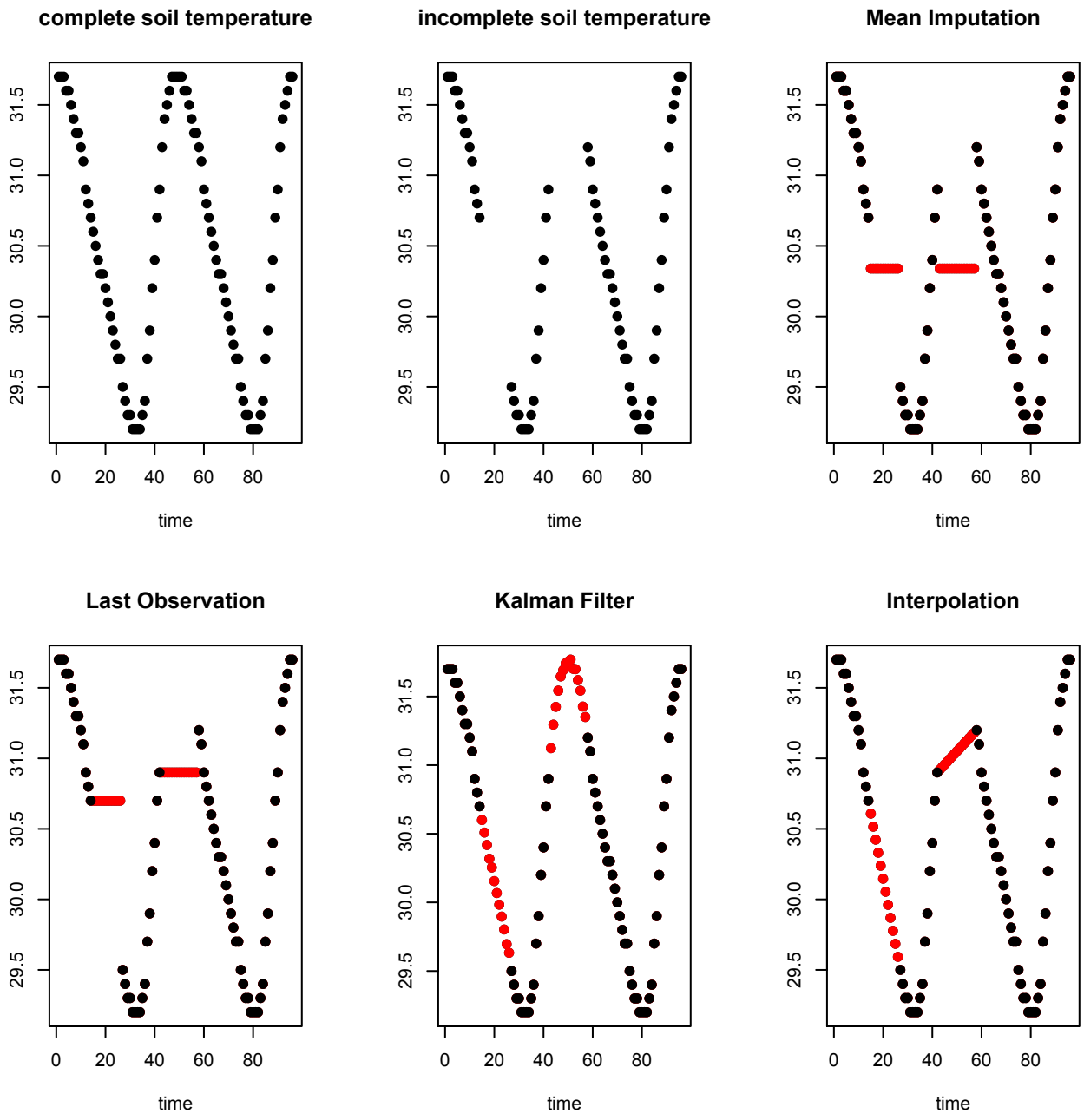
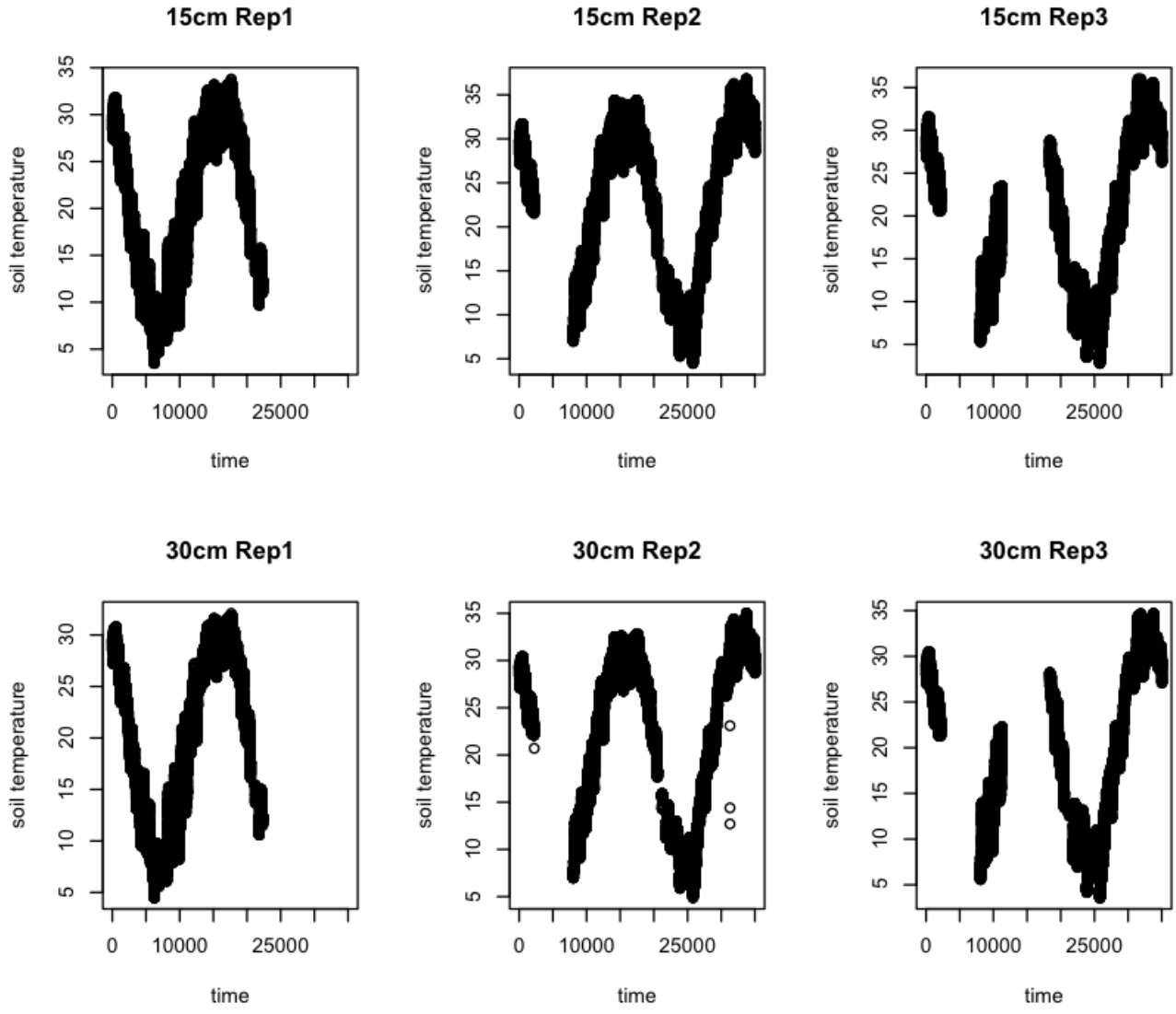
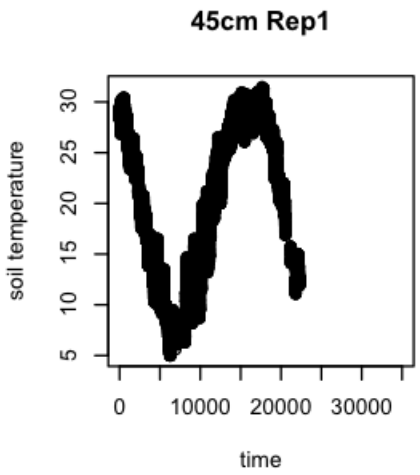
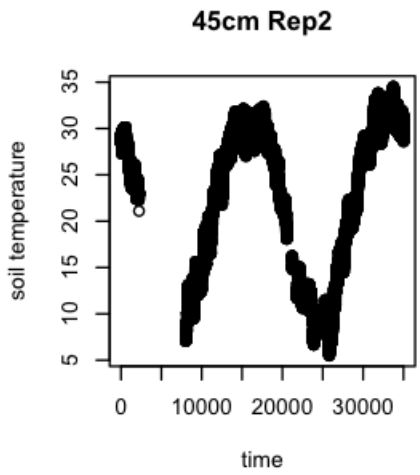
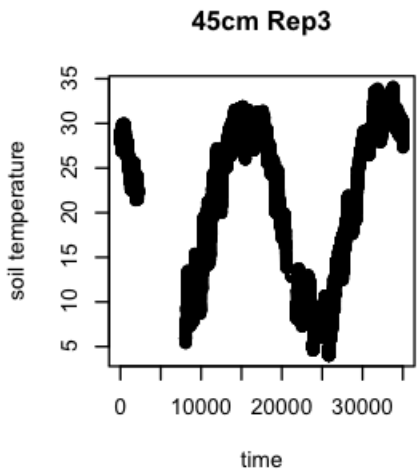


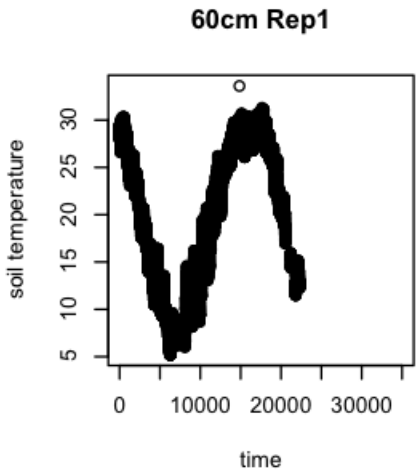
Figure 2.9: Plots of the Soil Temperature vs. Time for the 12 different combinations of Rep and distance.

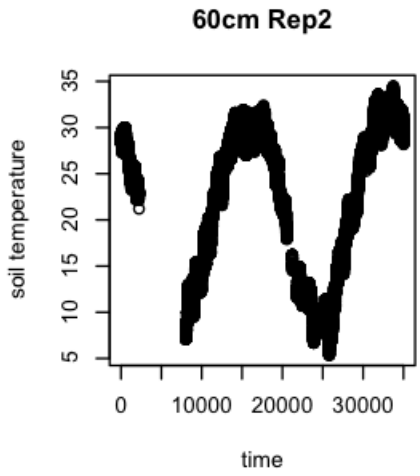



45cm Rep1
soil temperature
5
10
15
20
25
30
0
10000
20000
30000
time


45cm Rep2
soil temperature
5
10
15
20
25
30
35
0
10000
20000
30000
time


45cm Rep3
soil temperature
5
10
15
20
25
30
35
0
10000
20000
30000
time


60cm Rep1
soil temperature
5
10
15
20
25
30
0
10000
20000
30000
time


60cm Rep2
soil temperature
5
10
15
20
25
30
35
0
10000
20000
30000
time

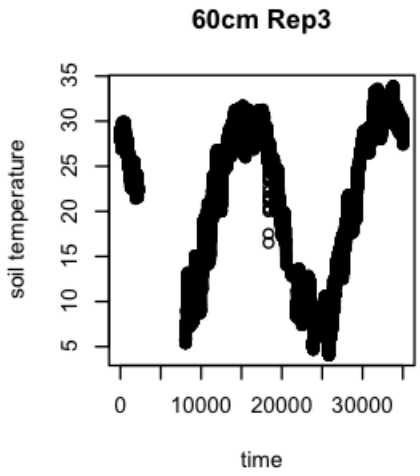
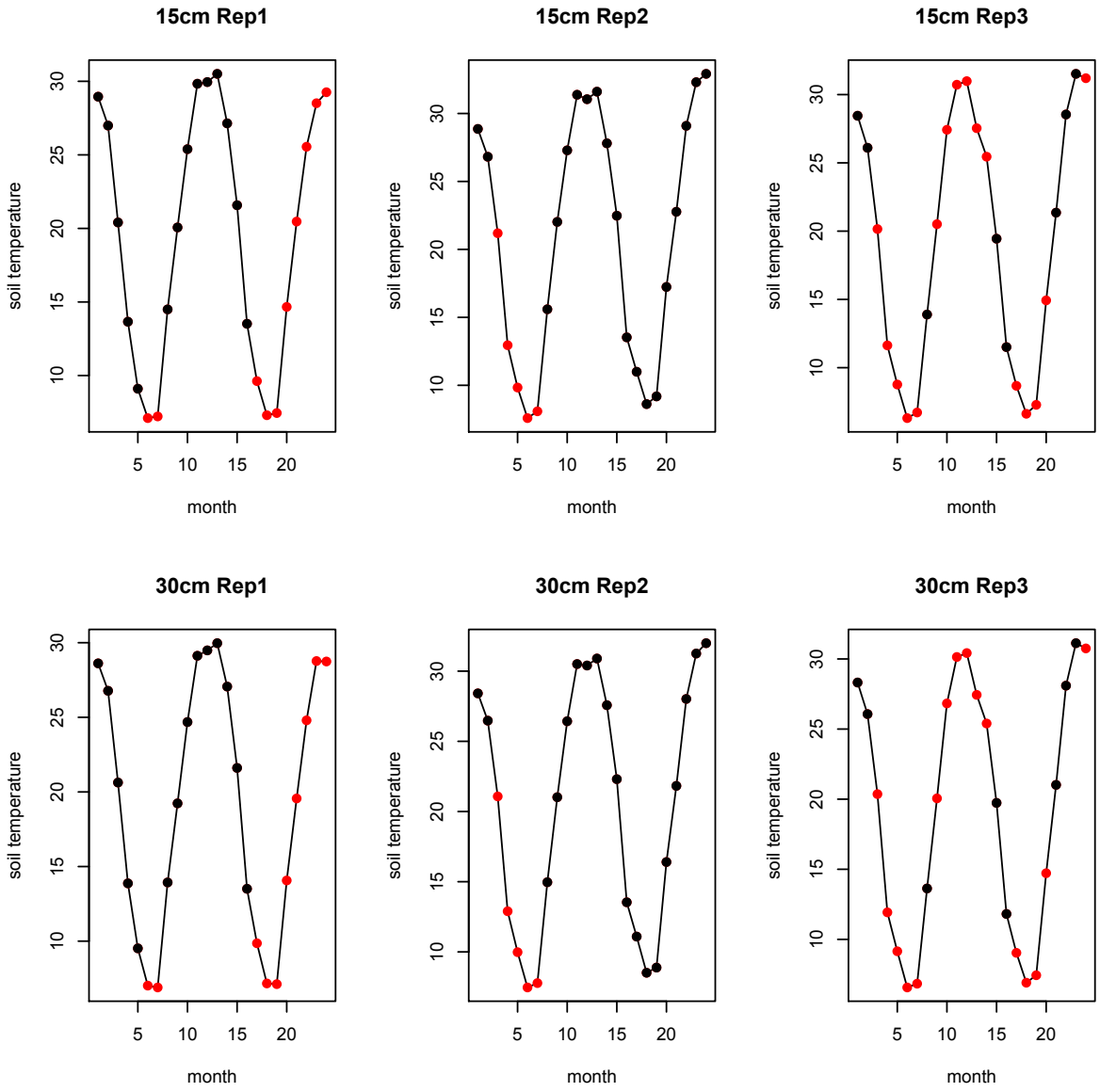
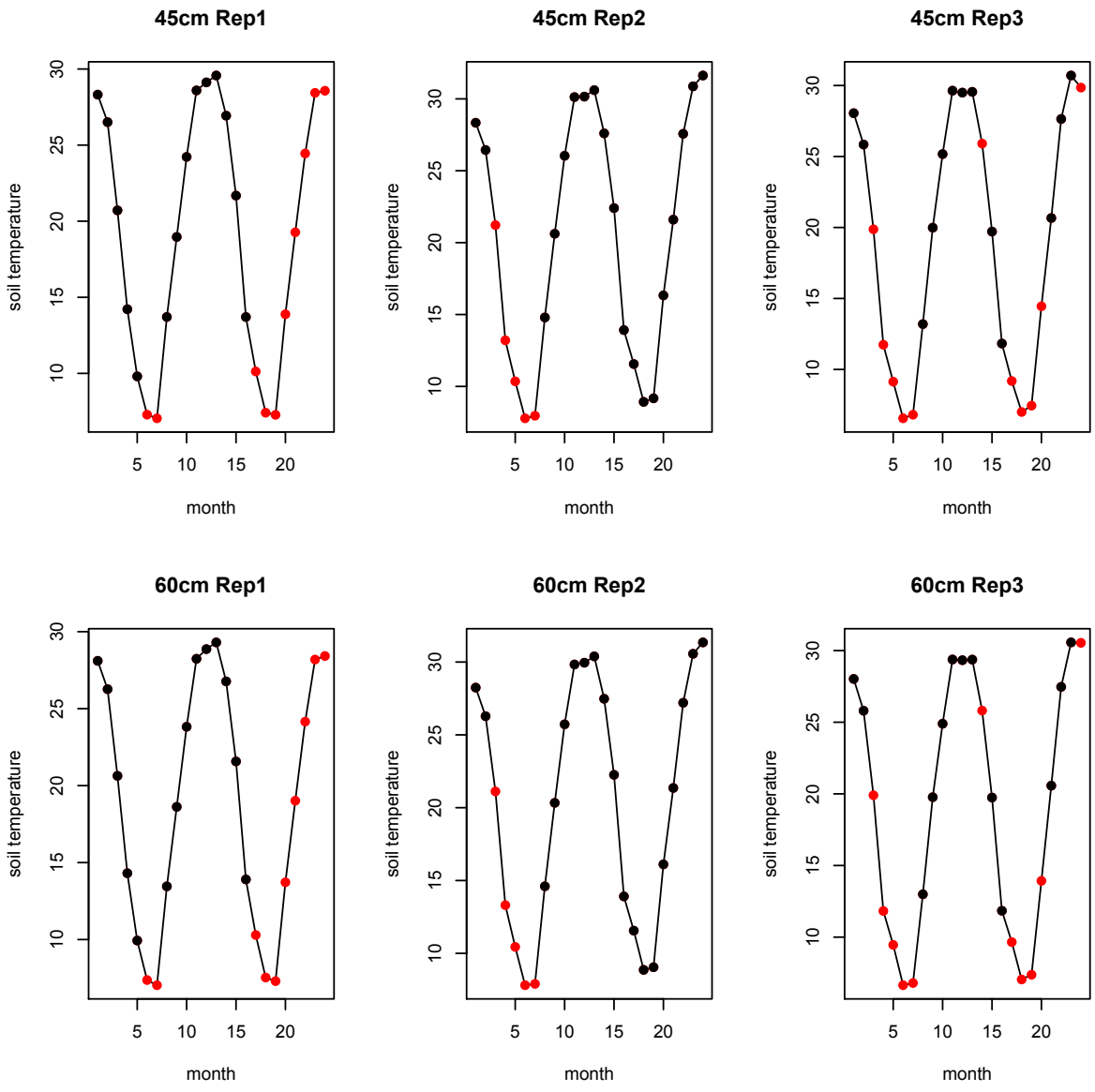

60cm Rep3
soil temperature
5
10
15
20
25
30
35
0
10000
20000
30000
time

Figure 2.10: Plot of the imputed Monthly Soil Temperature vs. Time for the 12 different combinations of Rep and distance with the actual data in black and the imputed data in red. Month 1 is September in 2013.




45cm Rep1
45cm Rep2
45cm Rep3
60cm Rep1
60cm Rep2
60cm Rep3
soil temperature
month
10
15
20
25
30
5
10
15
20

Chapter 3

Granger Causality Test

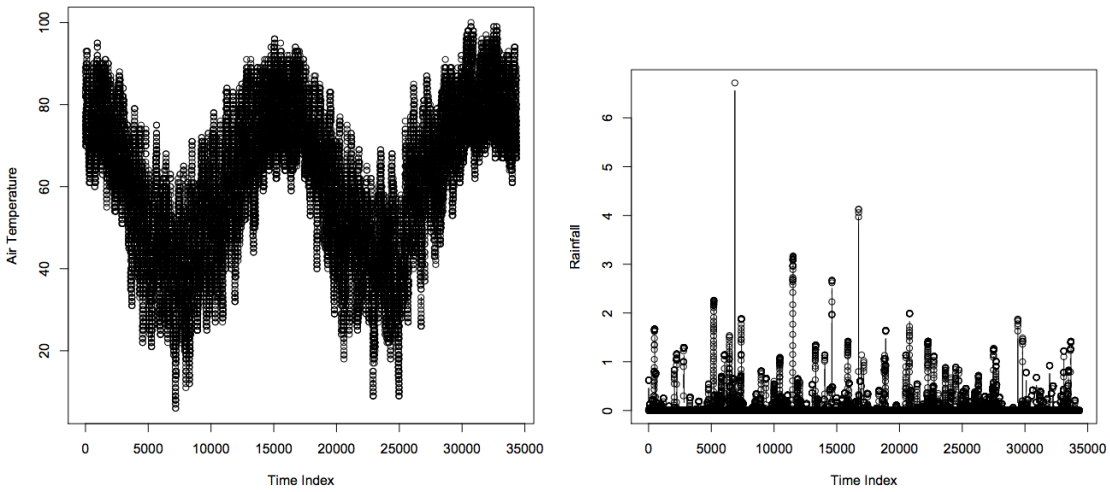
3.1 Introduction

This chapter focuses on the relationship between the soil temperature time series, and the air temperature, and rainfall time series. The air temperature and rainfall time series plots are shown in Figure 3.1. For air temperature (left panel), the time axis is the same scale as the soil temperature time axis since air temperature was recorded every 30 minutes for the 2-year period. The plot of air temperature shows seasonal pattern and shape similar to soil temperature. For rainfall (right panel), the time axis is also the same scale as soil temperature. The vertical axis is cumulative precipitation from the beginning of that day. The units of precipitation are millimetre. The plot show that rainfall is probably not seasonal. It is also important to note that there is only one air temperature time series and one rainfall time series (not 12 different series like soil temperature)

Because soil temperature, air temperature and rainfall are all time-series data, a Granger-Causality analysis was carried out in order to assess whether there is any potential predictability power of air temperature and rainfall for soil temperature. The conclusion that can be drawn is that air temperature and rainfall can be used to predict soil temperature, but the opposite, that the soil temperature can not be used to predict air temperature and rainfall.

"The Granger Causality test is a statistical hypothesis test for determining whether one time series somehow has a causal effect on another time series". It was proposed by Granger (1969). Ordinarily, regression coefficients are a measure of a relationship, but Granger argued that causality could be tested for by measuring the ability to predict the future values of a time series using prior

Figure 3.1: Plot of Air temperature and Rainfall



values of another time series. Since the question of "true causality" is very difficult to demonstrate in non-experimental data sets, and Granger assumed that one thing preceding another can be used as a proof of causation, econometricians assert that the Granger test finds only "predictive" or "Granger" causality" in (Diebold, 2001). In the Granger-sense, a previous value of x is a predictive cause of y if it is useful in forecasting a future y. In other words, previous x information is able to increase the accuracy of the forecast of y, as opposed to only using values of y.

3.2 Methodology Details

"A time series X is said to **Granger-cause** Y if it can be shown, usually through a series of t-tests and F-tests on lagged values of X, that those X values provide statistically significant information about future values of Y" (Kang, 1985).

There are 4 basic steps for running the Granger causality Test.

- 1. State the null hypothesis and alternative hypothesis in terms of model parameters. In our data set, the models define how air temperature or rainfall do or do not Granger-cause soil temperature. The two models are as follows

$$Y(t) = \sum_i \alpha_i Y(t-i) + e(t)$$

$$Y(t) = \sum_i \alpha_i Y(t-i) + \sum_j \beta_j X_1(t-j) + \sum_k \gamma_k X_2(t-k) + u(t) \tag{3.1}$$

The two equations in (3.1) represent a restricted model (top model, air temperature and rainfall do not Granger-cause soil temperature) and an unrestricted model (bottom model, air temperature and/or rainfall do Granger-cause soil temperature). The terms in the models include: $Y(t)$ is the soil temperature at time t, $X_1(t)$ is the air temperature at time t, $X_2(t)$ is the precipitation at time t, α, β and γ are coefficients, $e(t)$ and $u(t)$ are error terms. The errors terms are assumed to be independent.

Parameters from the two models in (3.1) can be used to test the two following hypotheses:

$$H_0 : \ \beta_j = 0, \ \ \gamma_k = 0 \ for\ all\ j\ and\ k \quad (restricted\ model)$$

$$H_1 : \ \beta_j \neq 0, \ \ \gamma_k \neq 0 \ for\ some\ j\ or\ k \quad (unrestricted\ model) \tag{3.2}$$

- 2. Choose the lags. One way to choose lags i, j and k is to run a model order test and another is to pick a series of lag values and run the Granger test several times. In either case, we chose AIC values as the criteria (Thornton & Batten, 1985). Among all these models, choose the one having the smallest AIC. And the lag lengths for $Y(t)$, $X_1(t)$ and $X_2(t)$ are not necessarily the same. The lag length for $Y(t)$ in the restricted and unrestricted models are the same.

- 3. Calculate the F-statistic and p-value from the model residuals in (3.1). The F-statistic equation is as follows:

$$F = \frac{(RSS_r - RSS_{ur})/(p_{ur} - p_r)}{RSS_{ur}/(n - p_{ur})} \sim F_{p_{ur}-p_r, n-p_{ur}} \tag{3.3}$$

$$RSS_{ur} = \sum_t \hat{u}(t), \qquad RSS_r = \sum_t \hat{e}(t)$$

where RSS_r, RSS_{ur} are sum of squared residuals of the restricted and unrestricted models,

respectively. n is the total number of observations, p_r and p_{ur} are the number of parameters in the restricted and unrestricted models respectively.

- 4. Reject null hypothesis if p-value is less than the significant level. In this dissertation, we choose 0.05 as the significant level.

3.3 Results

There are 35,040 observations for all these time series(after imputation). With this many observations, it was difficult to choose the reasonable number of lags to run the Granger-Causality test. Because our goal was to determine whether or not air temperature or rainfall had some influence on soil temperature, we decided to calculate the average soil temperature, air temperature, and rainfall for each month to reduce the number of observations. We then used this time series to choose lags and perform the Granger causality test. The dataset of mean monthly air temperature and rainfall are in Figure 3.2. Figure 3.2 indicates that the overall pattern of the original scale time series is similar to the pattern based on monthly averages. With the 2-year data set, there were 24 months. So there are at most 24 lags to use in the Granger-Causality test. From Chapter 1, recall that there were 3 reps (rep1, rep2 and rep3) and 4 different distances (15cm, 30cm, 45cm and 60cm), so there were 12 time-series of soil temperature. However, there was only one time-series for air temperature and only one for rainfall. Therefore the Granger-Causality test was conducted 12 times based on the 12 soil temperature time series.

The Granger Causality test results were calculated using PROC VARMAX with software SAS. The p-values are in Table 3.1. The results indicate air temperature and/or rainfall Granger-cause soil temperature at the 0.05 significance level in all 12 combinations. In other words, air temperature and rainfall were useful in predicting soil temperature in addition to previous soil temperature.

To determine whether air temperature only or rainfall only has an impact on soil temperature, we can use the following models to develop hypotheses and calculate the corresponding p-values:

Figure 3.2: Monthly average of Air temperature and Rainfall

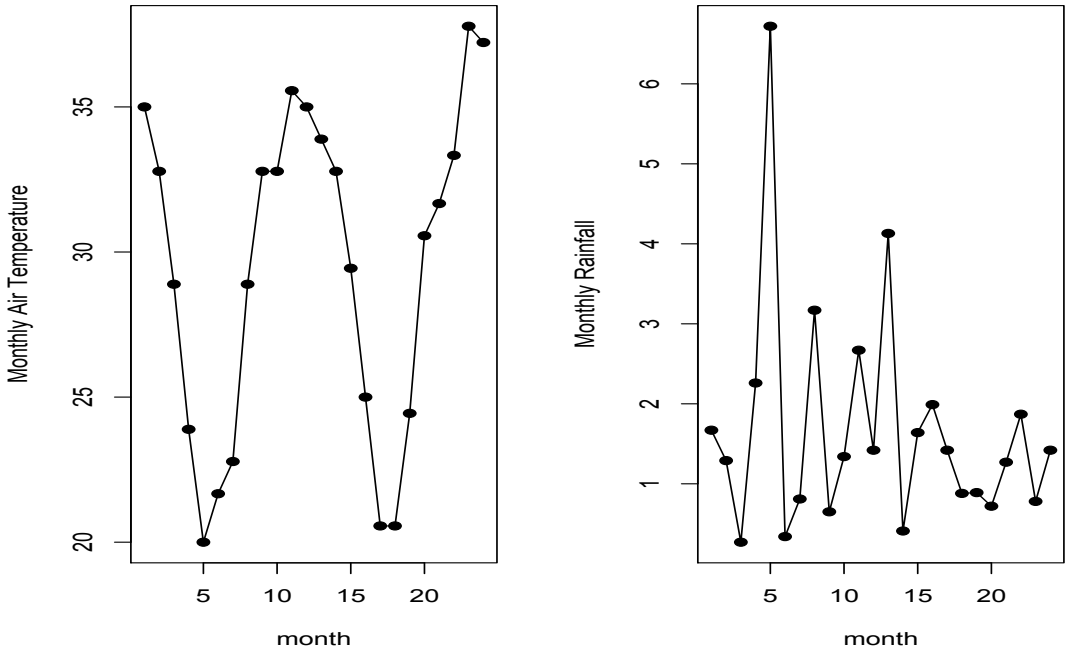


Table 3.1: Results of Granger Causality test

Restricted Model	Soil Temperature
Unrestricted Model	Air Temperature, Rainfall
P-value (Rep1, Dist15cm)	< 0.0001
P-value (Rep1, Dist30cm)	< 0.0001
P-value (Rep1, Dist45cm)	< 0.0001
P-value (Rep1, Dist60cm)	< 0.0001
P-value (Rep2, Dist15cm)	< 0.0001
P-value (Rep2, Dist30cm)	< 0.0001
P-value (Rep2, Dist45cm)	< 0.0001
P-value (Rep2, Dist60cm)	< 0.0001
P-value (Rep3, Dist15cm)	< 0.0001
P-value (Rep3, Dist30cm)	< 0.0001
P-value (Rep3, Dist45cm)	< 0.0001
P-value (Rep3, Dist60cm)	< 0.0001

Table 3.2: Results of Granger Causality test

Restricted Model	Soil Temperature
Unrestricted Model	Air Temperature
P-value (Rep1, Dist15cm)	< 0.0001
P-value (Rep1, Dist30cm)	< 0.0001
P-value (Rep1, Dist45cm)	< 0.0001
P-value (Rep1, Dist60cm)	< 0.0001
P-value (Rep2, Dist15cm)	< 0.0001
P-value (Rep2, Dist30cm)	< 0.0001
P-value (Rep2, Dist45cm)	< 0.0001
P-value (Rep2, Dist60cm)	< 0.0001
P-value (Rep3, Dist15cm)	< 0.0001
P-value (Rep3, Dist30cm)	< 0.0001
P-value (Rep3, Dist45cm)	< 0.0001
P-value (Rep3, Dist60cm)	< 0.0001

$$
\begin{aligned}
Y(t) &= \sum_i \alpha_i Y(t-i) + e(t) \\
Y(t) &= \sum_i \alpha_i Y(t-i) + \sum_j \beta_j X_1(t-j) + u(t)
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
Y(t) &= \sum_i \alpha_i Y(t-i) + e(t) \\
Y(t) &= \sum_i \alpha_i Y(t-i) + \sum_k \gamma_k X_2(t-k) + u(t)
\end{aligned} \tag{3.5}
$$

The hypotheses and models using (3.4) test the impact of predicting soil temperature by adding air temperature in addition to previous soil temperature; and the hypotheses and models using (3.5) test the impact of predicting soil temperature by adding rainfall in addition to previous soil temperature. The results in Table 3.2 indicate that air temperature does Granger-causes soil temperature at the 0.05 significance level, but the results in Table 3.3 indicate that rainfall does not Granger-cause soil temperature at the 0.05 significance level. Since some p-values from Table 3.3 are close to 0.05, there is still some possibility that rainfall Granger-causes soil temperature.

Table 3.3: Results of Granger Causality test

Restricted Model	Soil Temperature
Unrestricted Model	Rainfall
P-value (Rep1, Dist15cm)	0.0672
P-value (Rep1, Dist30cm)	0.0821
P-value (Rep1, Dist45cm)	0.0529
P-value (Rep1, Dist60cm)	0.0688
P-value (Rep2, Dist15cm)	0.1452
P-value (Rep2, Dist30cm)	0.0521
P-value (Rep2, Dist45cm)	0.1332
P-value (Rep2, Dist60cm)	0.1119
P-value (Rep3, Dist15cm)	0.0921
P-value (Rep3, Dist30cm)	0.0822
P-value (Rep3, Dist45cm)	0.0569
P-value (Rep3, Dist60cm)	0.0641

Chapter 4

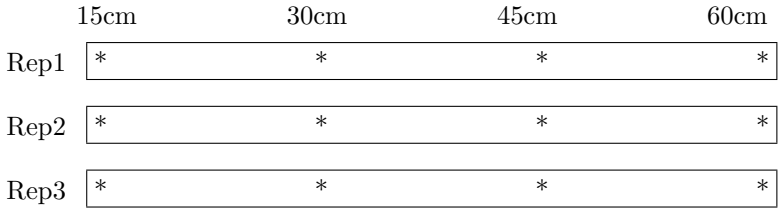
Experimental Design

4.1 Introduction

The overall objective of this dissertation was to determine the impact of sidewalks on soil temperature. Soil temperature was measured using sensors which were placed at 4 varying distances (15cm, 30cm, 45cm, 60cm) from the sidewalk. An important question was how exactly should the soil temperatures at these 4 distances be compared ? Designed experiments are one of the most common and useful statistical approaches which can be used to address this kind of questions. A designed experiment is characterized by the choice of treatments, the experimental units used, any covariates associated with the experimental units, the way treatments are allocated to the units, and how the responses are measured.

For this dataset, the treatments are the 4 distances from the sidewalk. The sites of the sensors are the experimental units. Possible covariates include air temperatures and rainfall. The sites are grouped within three different locations, reps, or blocks. Issues associated with "assigning" the distances to the sites will be discussed later. The measured response is the soil temperature time series. The treatments and the blocks are often denoted as factors in a designed experiment. The concept of fixed and random factors is also important in this study. With a fixed factor, all levels of the factor of interest are included in the experiment. With a random factor, all levels of interest are not in the experiment. In this soil temperature dataset, distance from the sidewalk (treatment) was considered as a fixed factor and location or rep (block) was considered as a random factor.

Figure 4.1: The design of the study: Rep or location is block, * is the Experimental Unit (E.U.) or the site of the sensor. We measure the response soil temperature, and two covariates (air temperature and rainfall) on the E.U.



The remainder of this chapter is organized as follows. Section 4.2 presents some basic concepts of randomized complete block designs. Section 4.3 provides the results of comparing the overall mean of the soil temperature time series at different distances. Results for comparing different characteristics of the soil temperature time series (other than the overall mean) at the different distances are also shown. The results are also adjusted for any impact of the air temperature, rainfall, and seasonality. Section 4.4 provides some introduction to a "peaks over threshold" approach to comparing the soil temperatures at different distance and results based on that approach are provided. Section 4.5 concludes with a summary discussion and opportunities for future research.

4.2 Randomized Complete Block Designs

The Randomized Complete Block Design (RCBD) is a commonly used design when similar experimental units are grouped into blocks or replicates. It is used to control variation in the experiment by accounting for the impact on the response due to factors such as location or time. With a randomized complete block design, experimental units are divided into groups called blocks, such that the variability within blocks is less than the variability among blocks (Oehlert, 2010). Then the experimental units within each block are assigned to treatments. Blocking serves many purposes. Within a block there is assumed homogeneity of experimental units, so treatment comparisons should be precise. Among blocks there is heterogeneity, so treatment comparisons are made across a wide variety of situations (Casella, 2008). As mentioned in the introduction, effects can be fixed or random. Blocks are typically considered as random factors. There is really no interest in the differences among the random blocks in the soil temperature study; they are used only because

Table 4.1: Data for a randomized complete block design

	Treatment				
Block	1	2	$\cdots$	t	Mean
1	Y_{11}	Y_{21}	$\cdots$	Y_{t1}	$\bar{Y}_{.1}$
2	Y_{12}	Y_{22}	$\cdots$	Y_{t2}	$\bar{Y}_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	Y_{1b}	Y_{2b}	$\cdots$	Y_{tb}	$\bar{Y}_{.b}$
Mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\cdots$	$\bar{Y}_{t.}$	$\bar{Y}_{..}$

the soil temperature may be different among the different blocks, and soil temperatures may be correlated within the blocks. Note that there is no treatment (distances) randomization in this design since the distances from the sidewalk cannot be randomly assigned in the reps. Including blocks as a random effect is a possible solution to the issue. The design is shown in Figure 4.1.

4.2.1 RCBD Theory

Consider RCBD similar to the soil temperature study as shown in Table 4.1. In Table 4.1, Y_{ij} is the observation for treatment i in block j; t is the number of treatments; b is the number of blocks; $\bar{Y}_{i.}$ is the sample mean for treatment i: $\bar{Y}_{i.} = \frac{1}{b}\sum_{j=1}^{b} Y_{ij}$; $\bar{Y}_{.j}$ is the sample mean for block j: $\bar{Y}_{.j} = \frac{1}{t}\sum_{i=1}^{t} Y_{ij}$; $\bar{Y}_{..}$ is the overall sample mean : $\bar{Y}_{..} = \frac{1}{tb}\sum_{i=1}^{t}\sum_{j=1}^{b} Y_{ij}$

The model for an observation in a randomized complete block design can be written in the form (Casella, 2008):

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, t;\ j = 1, \ldots, b, \tag{4.1}$$

where μ is an overall mean, τ_i are treatment effects, the error random variables $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ for $i = 1, \ldots, t$, and $j = 1, \ldots, b$, (normal errors with equal variances), the block effects $\beta_1, \ldots, \beta_b, \overset{iid}{\sim} N(0, \sigma_\beta^2)$ and are independent of ϵ_{ij} for all i, j.

Table 4.2: ANOVA table for RCBD

Source	SS	df	MS	F
Treatment	SST	t-1	MST=SST/(t-1)	MST/MSE
Block	SSB	b-1	MSB=SSB/(b-1)	
Error	SSE	(b-1)(t-1)	MSE=SSE/(b-1)(t-1)	
Total	TSS	bt-1		

The main goal in using the randomized complete block design was to examine the differences in the t treatment means. The null hypothesis does not have a difference among the treatment means while the alternative hypothesis treatment means do differ. That is:

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_t \quad \text{vs} \quad H_a: \ \mu_i\text{'s are not all equal.} \tag{4.2}$$

Analysis of variance (ANOVA) for RCBD can be used to test (4.2). The form of the ANOVA is shown in Table 4.2.

Note that the alternative hypothesis does not imply that all the μ_is are different, just that they are not all the same. We used ANOVA to test H_0 and found a small p-value (p=0.0005). So H_0 in (4.2) is rejected. Many options are available to determine how the treatment means differ (or determining the exact meaning of at least one μ_i differ from the rest). In this study, Fisher's Protected Least Significant Difference (LSD) Test was used to compare all pairs of soil temperature means from any two distances. For a specified significant level α, the least significant difference for comparing $\mu_i - \mu_j$ is:

$$LSD_{i,j} = t_{\alpha/2}\sqrt{s^2(1/n_i + 1/n_j)} \tag{4.3}$$

where α is the significant level, $t_{\alpha/2}$ is the critical t-value for area $\alpha/2$ and s^2 is the point estimator of σ^2 (the MSE from the ANOVA table). n_i and n_j are the respective sample sizes from treatment group i and j.

After calculating the LSD, we compared it to the difference in pairs of soil temperature

Table 4.3: Mean/Max/Min soil temperature (C) at different distances from the sidewalk. Means with the same letter in a column do not significantly differ based on ANOVA and Fisher's Protected (LSD) test with a significance level of 0.05.

Distance	Mean Soil Temp	Max Soil Temp	Min Soil Temp
15cm	19.94^a	25.65^a	15.27^a
30cm	19.66^a	23.94^{ab}	15.63^a
45cm	19.56^a	23.42^{ab}	15.97^a
60cm	19.47^a	23.13^b	16.01^a

sample means. This was used to test:

$$H_0 : \ \mu_i = \mu_j \ \ for\ i \neq j \tag{4.4}$$
$$H_1 : \ \mu_i \neq \mu_j \ \ for\ i \neq j$$

In the null hypothesis, the means of one pair of soil temperature are the same at a significant level 0.05. If $|Y_i - Y_j| \geq LSD_{i,j}$, declare the corresponding soil temperature means μ_i and μ_j are different. For each pairwise comparison of population means, the probability of a Type I error is fixed at a specified value of 0.05.

4.3 Results

4.3.1 Analysis based on a RCBD for the mean, max, and min of the soil temperature time series

Now we consider a RCBD for the overall mean of the soil temperature time series. The model is :

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \ \ i = 1, \ldots, 4; \ j = 1, \ldots, 3, \tag{4.5}$$

where Y_{ij} is the mean or max or min of the entire soil temperature time series, μ is an overall mean, τ_i are the effects of distance, β_j is the random effects of rep and ϵ_{ij} is error.

The results are shown in Table 4.3. Table 4.3 indicates that the highest sample mean and max soil temperature were both recorded at 15cm and the lowest sample mean and max soil

temperature were both recorded at 60cm. In other words, the closer to the sidewalk, the higher the sample mean and sample max of the soil temperature time series. However, the ANOVA and Fisher's Protected LSD test suggest that the true population means of the soil temperature time series at the different distances are not different (with a significance level of 0.05). The results suggest that the proximity to the sidewalk has no influence on the true mean soil temperature. While max soil temperature at 15cm are different from the soil temperature at 60cm. Max soil temperature at 30cm and 45cm are not significantly different. Max soil temperature at 45cm and 60cm are also not significantly different. For the min soil temperature, the closer to the sidewalk, the lower the min of the soil temperature which is opposite to the results for mean and max soil temperature. The results also suggest that the proximity to the sidewalk has no influence on the true min soil temperature with a significant level of 0.05.

4.3.2 Analysis based on a RCBD for mean, max, and min of the soil temperature time series by month

Now we consider the randomized complete block design (4.5) by month. In other words, we would compare characteristics of the soil temperature time series among the distances, based on the RCBD for each month. As in the previous section, we used different characteristics of the time series including the mean, minimum, and maximum soil temperature.

The results of mean, minimum, and maximum soil temperature are shown in Table 4.4, Table 4.5, and Table 4.6, respectively. Table 4.4-4.6 indicate that the closer to the sidewalk, the higher the sample mean, maximal or minimal soil temperature in warmer months (like May, June, July, August, Spetember). The colder months (like December, January, February) simply yield exact opposite results, the sample mean, maximal or minimal soil temperature is lowest when it is closest to the sidewalk. Depending on the different letters in one row (each month), the overall mean or minimum or maximum soil temperature for some pairs of different distances are significantly different with a significant level of 0.05 for that months.

Table 4.4: Mean soil temperatures (C) at different distances from the sidewalk for each month. Means with the same letter in a row do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05.

	Distance			
Month	15cm	30cm	45cm	60cm
$09-2013$	28.76^{a}	28.45^{b}	28.23^{bc}	28.12^{c}
$10-2013$	26.64^{a}	26.44^{ab}	26.26^{bc}	26.12^{c}
$11-2013$	20.58^{a}	20.69^{a}	20.60^{a}	20.55^{a}
$12-2013$	12.75^{a}	12.90^{a}	13.05^{a}	13.15^{a}
$01-2014$	9.23^{a}	9.55^{ab}	9.76^{b}	9.95^{c}
$02-2014$	6.99^{a}	7.02^{ab}	7.21^{ab}	7.28^{b}
$03-2014$	7.34^{a}	7.24^{a}	7.27^{a}	7.17^{a}
$04-2014$	14.66^{a}	14.18^{b}	13.90^{c}	13.68^{c}
$05-2014$	20.86^{a}	20.10^{b}	19.86^{b}	19.57^{b}
$06-2014$	26.70^{a}	25.98^{ab}	25.15^{bc}	24.81^{c}
$07-2014$	30.64^{a}	29.92^{b}	29.45^{c}	29.15^{d}
$08-2014$	30.66^{a}	30.10^{ab}	29.60^{bc}	29.38^{c}
$09-2014$	29.91^{a}	29.88^{a}	29.69^{a}	29.43^{a}
$10-2014$	26.79^{a}	26.81^{a}	26.68^{a}	26.67^{a}
$11-2014$	21.17^{a}	21.22^{a}	21.26^{a}	21.19^{a}
$12-2014$	12.85^{a}	12.95^{ab}	13.15^{b}	13.22^{b}
$01-2015$	9.76^{a}	9.99^{ab}	10.29^{bc}	10.50^{c}
$02-2015$	7.51^{a}	7.53^{a}	7.78^{b}	7.81^{b}
$03-2015$	7.97^{a}	7.97^{a}	7.91^{a}	13.50^{a}
$04-2015$	15.61^{a}	15.06^{b}	14.89^{bc}	14.59^{c}
$05-2015$	21.53^{a}	20.80^{b}	20.51^{bc}	20.32^{c}
$06-2015$	27.73^{a}	26.97^{b}	$26.55^{b}c$	26.28^{c}
$07-2015$	30.78^{a}	30.38^{ab}	30.00^{ab}	29.77^{b}
$08-2015$	31.12^{a}	30.49^{ab}	30.10^{b}	30.02^{b}

Table 4.5: Maximum soil temperatures (C) at different distances from the sidewalk for each month. Means with the same letter in a row do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05.

	Distance			
Month	15cm	30cm	45cm	60cm
$09-2013$	31.97^{a}	31.37^{ab}	31.37^{ab}	30.83^{b}
$10-2013$	31.67^{a}	30.57^{b}	30.17^{c}	30.07^{c}
$11-2013$	28.33^{a}	26.25^{ab}	25.61^{b}	25.25^{b}
$12-2013$	18.98^{a}	16.86^{a}	16.49^{a}	16.14^{a}
$01-2014$	18.45^{a}	14.93^{a}	14.13^{a}	13.61^{a}
$02-2014$	20.64^{a}	14.53^{b}	11.62^{c}	10.52^{c}
$03-2014$	20.99^{a}	15.59^{b}	13.73^{b}	12.90^{b}
$04-2014$	22.13^{a}	20.33^{b}	19.57^{c}	19.27^{c}
$05-2014$	27.03^{a}	25.73^{a}	25.97^{a}	25.53^{a}
$06-2014$	29.15^{a}	28.60^{a}	28.70^{a}	28.47^{a}
$07-2014$	32.26^{a}	31.59^{a}	31.30^{a}	31.21^{a}
$08-2014$	32.02^{a}	31.64^{a}	31.63^{a}	31.43^{a}
$09-2014$	31.81^{a}	31.29^{a}	31.77^{a}	31.67^{a}
$10-2014$	30.01^{a}	29.81^{a}	29.88^{a}	29.64^{a}
$11-2014$	26.77^{a}	26.200^{b}	25.76^{b}	25.71^{b}
$12-2014$	15.27^{a}	15.07^{a}	15.20^{a}	15.22^{a}
$01-2015$	16.32^{a}	14.01^{a}	13.45^{a}	13.17^{a}
$02-2015$	16.74^{a}	12.53^{a}	10.74^{a}	10.19^{a}
$03-2015$	18.11^{a}	14.44^{ab}	13.23^{ab}	12.67^{b}
$04-2015$	22.79^{a}	21.04^{ab}	20.26^{b}	19.90^{b}
$05-2015$	27.50^{a}	26.36^{a}	25.87^{a}	25.42^{a}
$06-2015$	29.98^{a}	29.68^{a}	29.52^{a}	29.17^{a}
$07-2015$	33.75^{a}	33.43^{a}	33.24^{a}	33.32^{a}
$08-2015$	32.91^{a}	32.90^{a}	32.76^{a}	32.81^{a}

Table 4.6: Minimum soil temperatures (C) at different distances from the sidewalk for each month. Means with the same letter in a row do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05.

	Distance			
Month	15cm	30cm	45cm	60cm
$09-2013$	24.23^{a}	24.03^{a}	23.97^{a}	23.93^{a}
$10-2013$	22.53^{a}	23.07^{b}	23.13^{b}	22.93^{b}
$11-2013$	16.29^{a}	16.20^{b}	15.93^{b}	15.01^{b}
$12-2013$	9.00^{a}	9.85^{b}	10.32^{bc}	10.50^{c}
$01-2014$	4.46^{a}	5.31^{b}	5.84^{c}	6.17^{c}
$02-2014$	3.49^{a}	4.28^{b}	5.07^{c}	5.36^{c}
$03-2014$	2.24^{a}	2.98^{b}	3.34^{b}	3.42^{b}
$04-2014$	8.83^{a}	9.50^{b}	9.70^{b}	9.77^{b}
$05-2014$	15.02^{a}	15.32^{a}	14.47^{a}	14.30^{a}
$06-2014$	21.75^{a}	21.98^{a}	20.61^{a}	20.50^{a}
$07-2014$	26.08^{a}	25.87^{a}	26.17^{a}	25.93^{a}
$08-2014$	26.90^{a}	26.36^{a}	26.36^{a}	26.14^{a}
$09-2014$	25.66^{a}	25.85^{a}	27.17^{a}	27.03^{a}
$10-2014$	21.87^{a}	22.53^{b}	22.83^{b}	22.72^{b}
$11-2014$	14.70^{a}	15.73^{b}	16.27^{b}	16.30^{b}
$12-2014$	9.03^{a}	9.83^{b}	10.30^{bc}	10.47^{c}
$01-2015$	4.86^{a}	5.63^{ab}	6.31^{bc}	6.78^{c}
$02-2015$	4.03^{a}	4.77^{b}	5.61^{c}	5.86^{c}
$03-2015$	3.06^{a}	3.68^{b}	4.27^{c}	4.38^{c}
$04-2015$	10.43^{a}	10.71^{a}	10.94^{a}	11.08^{a}
$05-2015$	16.67^{a}	16.85^{a}	16.99^{a}	16.93^{a}
$06-2015$	23.94^{a}	23.84^{a}	23.47^{a}	23.18^{a}
$07-2015$	26.77^{a}	26.17^{a}	26.78^{a}	26.66^{a}
$08-2015$	26.95^{a}	27.65^{a}	27.97^{a}	27.13^{a}

Table 4.7: Mean/Max/Min soil temperature (C) at different distances from the sidewalk using model (4.6). Means with the same letter in a column do not significantly differ based on ANOVA and Fisher's Protected (LSD) test with a significance level of 0.05.

Distance	Mean Soil Temp	Max Soil Temp	Min Soil Temp
15cm	20.23^{a}	25.72^{a}	15.31^{a}
30cm	19.78^{b}	25.13^{b}	15.59^{a}
45cm	19.62^{bc}	24.82^{bc}	16.03^{b}
60cm	19.49^{c}	24.53^{c}	16.42^{c}

4.3.3 Analysis based on a RCBD for the mean, max, min of the soil temperature time series with month included in the model

We applied RCBD to our soil temperature dataset. There are 4 treatments (15cm, 30cm, 45cm, 60cm) and 3 blocks (rep1, rep2, rep3). For each distance and each rep, there are 24 monthly soil temperatures (maxima, average and minimal). So we typically take 24 measurements on each E.U. over time.

We used the following model:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{aij} + T_k + \tau T_{ik} + \epsilon_{bijk}; \;\; i = 1, 2, 3, 4; \;\; j = 1, 2, 3; \;\; k = 1, 2, \ldots, 24; \qquad (4.6)$$

where Y_{ijk} is soil temperature (max, mean or min), μ is an overall mean, τ_i is effect of distance, β_j is the random effect of rep, T_k is the effect of time (month).ϵ_{aij} is the random effect of block j in treatment i, ϵ_{bijk} is the experimental random error.

The results for soil temperature with maximal, average, and minimal as response are displayed in Figure 4.2, Figure 4.3, and Figure 4.4. Table 4.7 shows the mean, max, and min soil temperature for increasing distances from the sidewalk with model (4.6). Table 4.7 indicates that the highest sample mean and max soil temperature were both recorded at 15cm and the lowest sample mean and max soil temperature were both recorded at 60cm. However, the closer to the sidewalk, the lower the sample min of the soil temperature time series. According to the ANOVA and Fisher's Protected LSD test, the true population mean, max, and min of the soil temperature time series at the different distances are different (with a significance level of 0.05). The results suggest that the proximity to the sidewalk has influence on the true mean, max, and min soil temperature.

Figure 4.2: Mean soil temperatures (C) at different distances (15^a cm, 30^b cm, 45^{bc} cm, 60^c cm) from the sidewalk for each month. Means with the same letter do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05

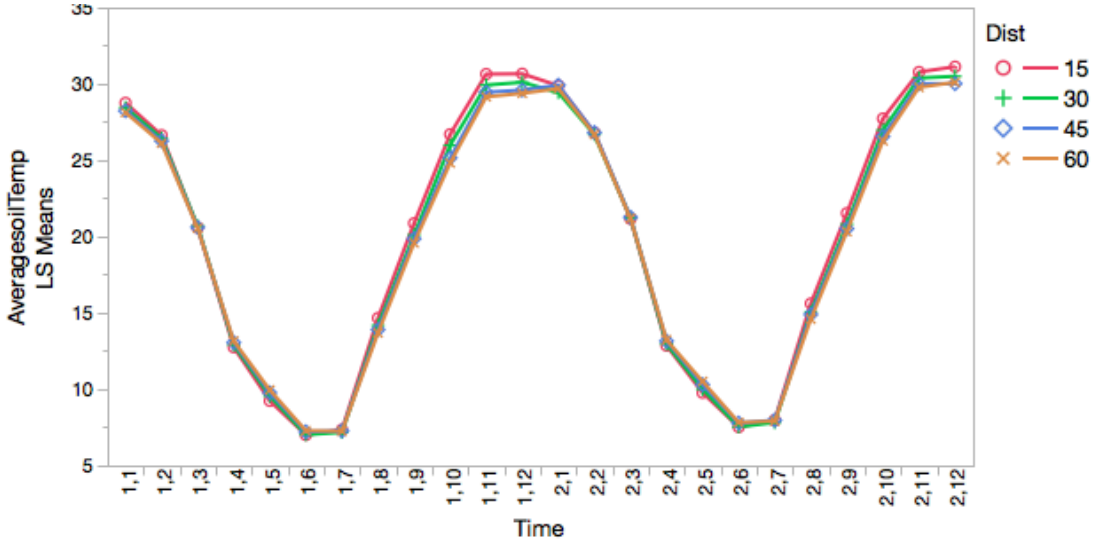


Figure 4.3: Max soil temperatures (C) at different distances (15^a cm, 30^b cm, 45^{bc} cm, 60^c cm) from the sidewalk for each month. Means with the same letter do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05

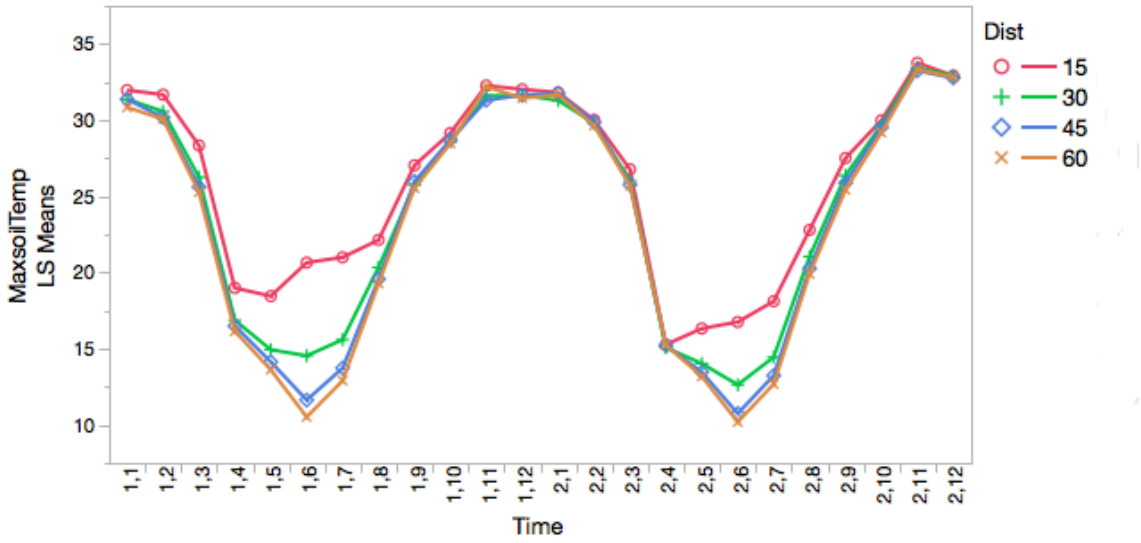
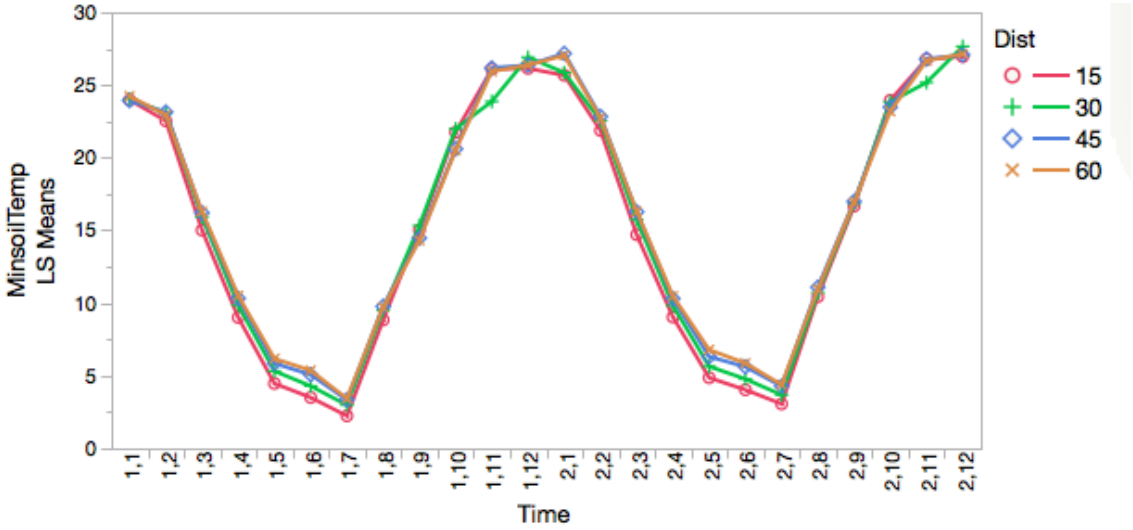


Figure 4.4: Min soil temperatures (C) at different distances (15^a cm, 30^b cm, 45^{bc} cm, 60^c cm) from the sidewalk for each month. Means with the same letter do not significantly differ based on ANOVA and Fisher's Protected LSD test with a significance level of 0.05



4.3.4 Analysis based on a RCBD for mean, max, min soil temperature time series with air temperature included in the model

From Chapter 3, we know that air temperature have an influence on soil temperature. So soil temperature (measured response) is related not only to the distance (treatment) but also to air temperature (covariate). And the distances do not have an effect on air temperature. In experiment design, covariate is defined as the variables that describe the differences in experimental units or experimental conditions (Ott & Longnecker, 2010). The issue is that the covariates (air temperature) are all the same for different experimental units (sites of sensor). A plot of the soil temperature for each distance is shown in Figure 4.5 with the covariate, air temperature given on the horizontal axis.

Figure 4.5 indicated that the relationship between soil temperature and the covariate air temperature is almost linear or curvilinear. Since the analysis of covariates combines features of the analysis of variance and regression analysis, we make use of a general linear model formulation for the analysis of the soil temperature time series data. If we assume a linear relationship between soil temperature Y_{ijk} and the covariate, air temperature x_{ijk} for each treatment, we have a RCBD with 3 blocks (rep1, rep2, rep3), 4 treatments (15cm, 30c, 45c, 60cm), one covariate (air temperature) and $n = 24$ observations per treatment in each block. In section 4.3.3, we add the effect of time (month) in the RCBD model (4.6). However, the effect of month is not included in the following

Figure 4.5: Soil temperature for 4 distances with covariate, air temperature

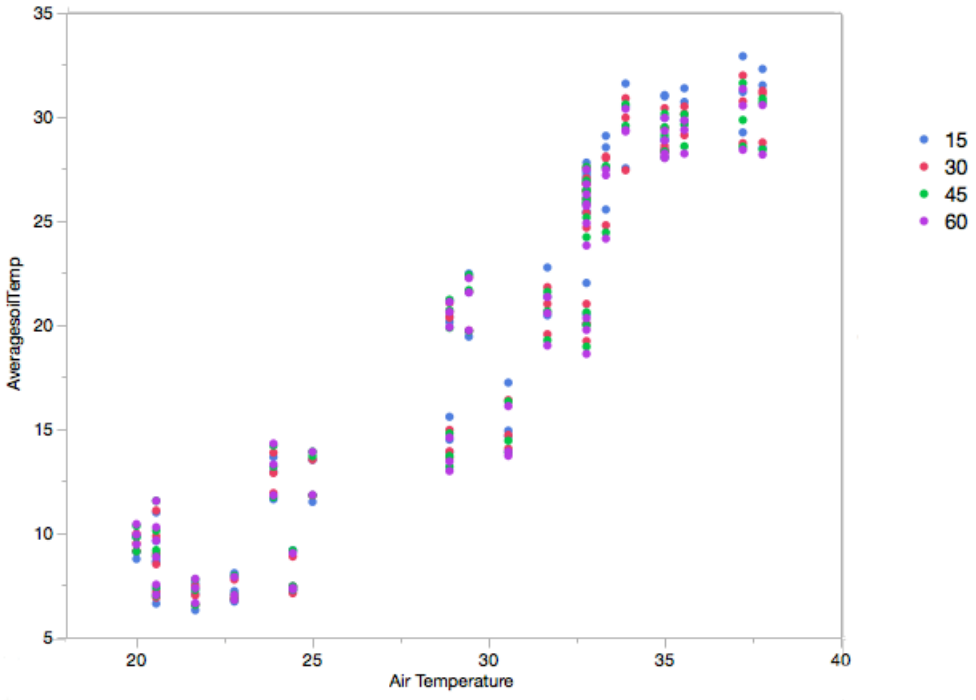


Table 4.8: Expected values for the RCBD with one covariate

Block	Treatment	Expected Responses
1	1	$\beta_0 + \beta_1 x_1$
1	2	$(\beta_0 + \beta_4) + (\beta_1 + \beta_7)x_1$
1	3	$(\beta_0 + \beta_5) + (\beta_1 + \beta_8)x_1$
1	4	$(\beta_0 + \beta_6) + (\beta_1 + \beta_9)x_1$
2	1	$(\beta_0 + \beta_2) + \beta_1 x_1$
2	2	$(\beta_0 + \beta_2 + \beta_4) + (\beta_1 + \beta_7)x_1$
2	3	$(\beta_0 + \beta_2 + \beta_5) + (\beta_1 + \beta_8)x_1$
2	4	$(\beta_0 + \beta_2 + \beta_6) + (\beta_1 + \beta_9)x_1$
3	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
3	2	$(\beta_0 + \beta_3 + \beta_4) + (\beta_1 + \beta_7)x_1$
3	3	$(\beta_0 + \beta_3 + \beta_5) + (\beta_1 + \beta_8)x_1$
3	4	$(\beta_0 + \beta_3 + \beta_6) + (\beta_1 + \beta_9)x_1$

model which contains air temperature. Because the monthly air temperature is the same for any reps and any distances.

$$Y_{ijk} = \beta_0 + \tau_i + b_j + \delta_i x_{ijk} + \epsilon_{ijk} \tag{4.7}$$

where $i = 1, 2, 3, 4$; $j = 1, 2, 3$; and $k = 1, \cdots, 24$. Y_{ijk} is the mean or max or min of the monthly soil temperature, x_{ijk} is the mean of the monthly air temperature, β_0 is the intercept of the regression of y on x, τ_i is the ith treatment effect, δ_i is the slope of the regression of y on x, b_j is the jth random block effect, and ϵ_{ijk}s are the random errors. We can write this in a general linear model as:

$$\begin{aligned} \text{Full Model}: \quad y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \\ & + \beta_7 x_1 x_4 + \beta_8 x_1 x_5 + \beta_9 x_1 x_6 + \epsilon_{ijk} \end{aligned} \tag{4.8}$$

where x_1 = covariate; $x_2 = 1$ if block 2 , $x_2 = 0$ otherwise; $x_3 = 1$ if block 3 , $x_3 = 0$ otherwise; $x_4 = 1$ if treatment 2, $x_4 = 0$ otherwise; $x_5 = 1$ if treatment 3, $x_5 = 0$ otherwise; $x_6 = 1$ if treatment 4 , $x_6 = 0$ otherwise.

In the full model (4.8), the response y is related to a quantitative variable x_1 and two qualitative variables: blocks and treatments. An interpretation of βs in the model is shown in Table 4.8:

The model (4.8) provides for a linear relationship between y and x_1 for each of the treatments in each block, and it also allows for differences among intercepts and slopes. Note that the treatments

have different slopes, but that each treatment has the same slope across blocks (Ott & Longnecker, 2010). To test whether the linear relationship between soil temperature and air temperature is the same for the 4 distances (whether the 4 lines have equal slopes), we fit a model to the soil temperature data in which the 4 lines have the same slope, but different intercepts.

$$\text{Reduced Model I} : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon_{ijk} \tag{4.9}$$

A test for equal slopes is obtained by testing in Reduced Model I the hypotheses

$$H_0 : \ \beta_7 = \beta_8 = \beta_9 = 0 \quad \text{vs} \quad H_a : \ \beta_i\text{'s are not all equal}, i = 7, 8, 9$$

The test F-statistic for H_0 versus H_1 is

$$F = \frac{(SSE_{RI} - SSE_F)/(df_R - df_F)}{SSE_F/df_F} \tag{4.10}$$

where SSE_{RI} is the sum of squares error from the Reduced Model I, SSE_F is the sum of squares error from the Full Model. Based on this F-statistic, we calculated a p-value. We used mean, max, min soil temperature as the response, respectively. We all calculated a p-value which is greater than the significant level 0.05. Thus, the 4 treatments have the same slope. Then, we can test for differences among the 4 treatments. We would fit a model without treatment differences:

$$\text{Reduced Model II} : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_{ijk}$$

Then the following null hypothesis can be used to test for differences among treatments.

$$H_0 : \ \beta_4 = \beta_5 = \beta_6 = 0 \quad \text{vs} \quad H_a : \ \beta_i\text{'s are not all equal}, i = 4, 5, 6$$

The test F-statistic for H_0 versus H_1 is

$$F = \frac{(SSE_{RII} - SSE_{RI})/(df_{RII} - df_{RI})}{SSE_{RI}/df_{RI}} \tag{4.11}$$

where SSE_{RII} is the sum of squares error from the Reduced Model II. Based on this F-statistic, we

Table 4.9: Mean/Max/Min soil temperature (C) at different distances from the sidewalk with the impact of air temperature. Means with the same letter in a column do not significantly differ with a significance level of 0.05.

Distance	Mean Soil Temp	Max Soil Temp	Min Soil Temp
15cm	20.11^{a}	24.98^{a}	14.89^{a}
30cm	19.77^{ab}	24.25^{b}	15.34^{b}
45cm	19.66^{ab}	23.79^{bc}	15.68^{bc}
60cm	19.35^{b}	23.52^{c}	15.92^{c}

can calculated the p-value. When the p-value is less than the significant level, we then used Fisher's Protected LSD test to compare any pair of soil temperature. The results are shown in Table 4.9.

4.3.5 Analysis based on a RCBD for residuals of the mean, max, min soil temperature time series

Another approach to analyze the soil temperature with the influence of air temperature is to use the residual from the function:

$$Y = f(x) \tag{4.12}$$

Where f is a function to describe the relationship between soil temperature and air temperature. We use mean air temperature as the input x and maximal, average or minimal soil temperature as the output Y in (4.12).

We substract the effect of air temperature from soil temperature, then use the difference or residual $(y - f(x))$ as the response in model (4.6). In this study, we used linear model and Granger-causality model as f, respectively. Then we used the residual from these two models as the response in model (4.6), respectively.

The results for residuals from linear model and Granger-causality model are in Table 4.10 and Table 4.11. From the results, we can conclude that the true residuals from the linear model are significantly different among incremental distances (15cm, 30cm, 45cm, 60cm) from the sidewalk. While the true residuals from the Granger-causality model do not significantly differ among incremental distances (15cm, 30cm, 45cm, 60cm) from the sidewalk. When average and max monthly soil temperature are used as output in (4.12), the highest residual from linear model and Granger-causality model was both recorded at 15 cm and the lowest was both recorded at 60 cm. When min monthly soil temperature are used as output in (4.12), the closer to the sidewalk the lower the

Table 4.10: Variation of residual from linear model for increasing distances from the sidewalk. Means with the same letter in a column do not significantly differ with a significance level of 0.05.

Distance	Mean residual	Max residual	Min residual
15cm	0.284^{a}	1.616^{a}	-0.415^{a}
30cm	0.004^{b}	-0.092^{b}	-0.189^{ab}
45cm	-0.093^{bc}	-0.619^{bc}	0.287^{b}
60cm	-0.195^{c}	-0.905^{c}	0.317^{b}

Table 4.11: Variation of residual from Granger Causality model for increasing distances from the sidewalk. Means with the same letter in a column do not significantly differ with a significance level of 0.05.

Distance	Mean residual	Max residual	Min residual
15cm	0.115^{a}	0.132^{a}	-0.075^{a}
30cm	0.011^{a}	0.122^{b}	0.018^{a}
45cm	-0.041^{a}	0.018^{bc}	0.012^{a}
60cm	-0.085^{a}	-0.015^{c}	0.107^{a}

residual for both models.

4.4 Peaks Over Threshold

The scope of this section is to analyze temperatures using a Peaks Over Threshold (POT) approach. The purpose of this POT approach is to analyze soil temperature extremes by comparing the number of soil temperature that exceed a high threshold for different distances. This is motivated by the fact that in the analysis, the max soil temperature seems to be the time series characteristics that was influenced most by the sidewalk distance.

First, we need to choose the threshold. Commonly used procedure consists of choosing one of the sample points as a threshold, the choice is often the kth upper order statistic Y_{n-k+1} from the ordered sequence $Y_1, \cdots, Y_n$ (Bommier, 2014). Frequently used is the 0.90 quantile. We first calculate the 0.90 quantile for the daily soil temperature for each rep and each month. The thresholds are shown in Figure 4.6. We then count the number of observations exceeding thresholds and use these values as the response in model (4.6). The results are in Table 4.11

Figure 4.6: The thresholds for different months and different reps

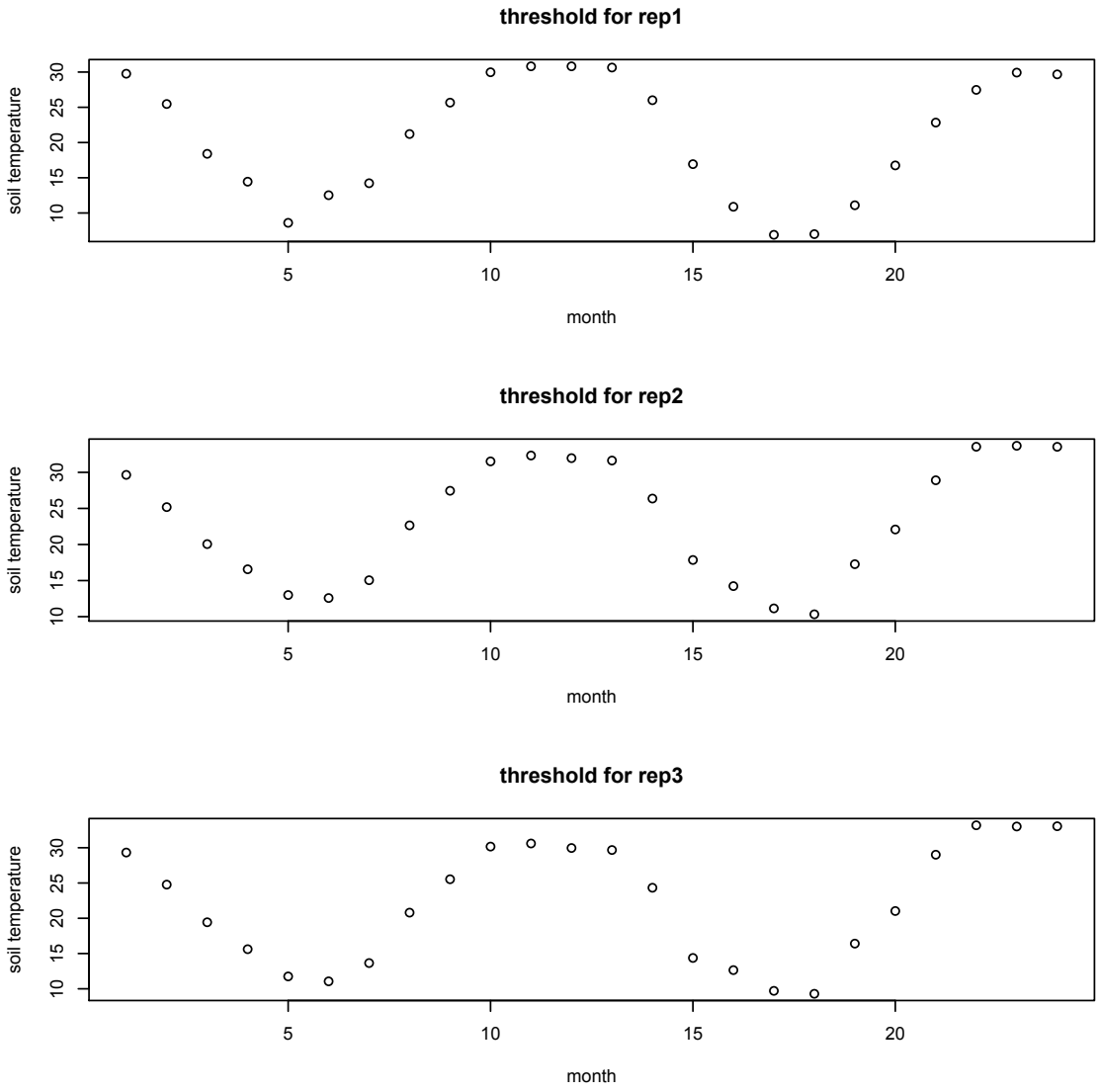


Table 4.12: Average number of exceedances for increasing distances from the sidewalk by month. Means with the same letter in a row do not significantly differ with a significance level of 0.05.

	Average number of exceedances			
Month	15cm	30cm	45cm	60cm
$09-2013$	8^{a}	3.67^{b}	0.33^{c}	0^{c}
$10-2013$	5.67^{a}	4^{b}	2.33^{c}	1^{c}
$11-2013$	4.67^{a}	4^{a}	1.67^{a}	1.67^{a}
$12-2013$	5.67^{a}	5.33^{a}	1^{b}	1^{b}
$01-2014$	5.33^{a}	5^{a}	1.33^{a}	1.33^{a}
$02-2014$	6^{a}	2.67^{b}	2^{b}	1.33^{b}
$03-2014$	7.33^{a}	3.67^{b}	1.33^{c}	0.67^{c}
$04-2014$	5.33^{a}	3.67^{b}	2.33^{b}	0.67^{c}
$05-2014$	9.33^{a}	3.33^{b}	0.33^{c}	14.30^{0}
$06-2014$	9^{a}	3^{b}	0^{c}	0^{c}
$07-2014$	11.33^{a}	1.67^{b}	0^{b}	0^{b}
$08-2014$	9^{a}	4^{b}	0^{c}	0^{c}
$09-2014$	7^{a}	4.33^{b}	0.67^{c}	0^{c}
$10-2014$	5^{a}	3^{a}	3^{a}	2^{a}
$11-2014$	3.67^{a}	3.33^{a}	2.67^{a}	2.33^{a}
$12-2014$	4.33^{a}	4^{a}	2.67^{ab}	2^{b}
$01-2015$	5.33^{a}	3.67^{a}	2.33^{a}	1.67^{a}
$02-2015$	4.33^{a}	4.33^{a}	2^{b}	1.33^{b}
$03-2015$	6.67^{a}	3^{b}	2^{b}	1.33^{b}
$04-2015$	8^{a}	3^{b}	0.67^{c}	0.33^{c}
$05-2015$	9.67^{a}	3^{b}	0.33^{c}	0^{c}
$06-2015$	9.67^{a}	2^{b}	0.33^{b}	0^{b}
$07-2015$	11.67^{a}	1.33^{b}	0^{b}	0^{b}
$08-2015$	8.33^{a}	3.67^{ab}	1^{b}	0^{b}

4.5 Discussion and Future Work

This work has developed a complete block design based on the monthly soil temperature time series data with the impact of air temperature. For future research, we will consider to build a more general design for the original data set (sample size is 35,040). And we will consider some other methods to add the influence of air temperature into our design. In addition, we can consider spatial space model. Since the soil temperatures were measured over distances (15cm, 30cm, 45cm, 60cm), it is reasonable to assume that the correlation between any two observations decrease as the sensor are further apart in distance.

Bibliography

[1] Bommier, E. (2014). Peaks-over-threshold modelling of environmental data.

[2] Casella, G. (2008). Statistical design. Springer Science & Business Media.

[3] D. Little, R.J.,A & Rublin, D.B.(1987). Statistical Analysis with Missing Data, New York, NY: Wiley.

[4] Diebold, F. X. (2001). *Elements of Forecasting (2nd ed.).* Cincinnati (Ohio: South-Western, Thomson learning.

[5] Dixon, W. J. (1988), BMDP Statistical Software Manual to Accompany the... Software Release. University of California Press.

[6] Drennan, K., Wijesinghe, D., White, S., Menchyk, N., Vincent, E., & Park, D. (2014). Soil Temperature and Moisture Differentials from Concrete Sidewalks.

[7] Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods* (Vol. 38). Oxford University Press.

[8] Enders, C. K. (2010), *Applied Missing Data Analysis.* The Guilford Press.

[9] Ersavas, B. F. (2014). U.S. Patent No. 8,649,907. Washington, DC: U.S. Patent and Trademark Office.

[10] Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete data in sample surveys,* 2(Part IV):185-207.

[11] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral Methods. *Econometrica: Journal of the Econometric Society,* 424438.

[12] Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter.* Cambridge university press.

[13] Hyndman, R. J.(2014). *forecast: Forecasting functions for time series and linear models.* Retrieved from http://CRAN.R- project.org/package=forecast. R package version 5.5.

[14] Jamshidian, M., Jalal, S. J., & Jansen, C. (2014). MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR).

[15] Kang, H. (1985). The effects of detrending in Granger causality tests. *Journal of Business & Economic Statistics, 3*(4), 344-349.

[16] Little, R. J. (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 11981202.

[17] Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. Sociological Methods & Research, 18(2-3), 292-326.

[18] Moritz, S., & Bartz-Beielstein, T (2015). imputeTS: time series missing value imputation in R. R package version 0.4.

[19] Moritz, S., Sard, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924*.

[20] Oehlert, G. W. (2010). A first course in design and analysis of experiments.

[21] Ott, L., & Longnecker, M. (2010). An introduction to statistical methods and data analysis (6th ed.). Belmont, CA: Brooks/Cole.

[22] Rubin, D. B. (1976) Inference and missing data. *Biometrika*, 63(3):581592.

[23] Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys*(Vol 17). Wiley.

[24] Samarin, M. (2012), Linear Interpolation, *Encyclopedia of Mathematics*. Retrieved from http://www.encyclopediaofmath.org/index.php/Linear interpolation/oldid=27068.

[25] Satterthwaite, D., McGranahan, G., & Tacoli, C. (2010). Urbanization and its implications for food and farming. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 365 (1554), 2809-2820.

[26] Shi, B., Shao, Y., Liu, C., & Wang, B., (2009). Impact and key issues of urban heat island effect to soil engineering properties[J]. *Journal of Engineering Geology*, 2, 006.

[27] Swanson, D. A., Tayman, J., & Bryan T. M. (2011). MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts.*Journal of Population Research, 28*(2-3), 225-243.

[28] Thornton, D. L., & Batten, D. S. (1985). Lag-length selection and tests of Granger causality between money and income. Journal of Money, credit and Banking, 17(2), 164-178.

[29] Vacek, P.M., & Ashikaga, T (1980). An examination of the nearest neighbor rule for imputing missing values.*Proc. Statist. Computing Sect., Amer. Statist. Ass,* 326-331.

[30] Welch, B. L. (1974). The generalization of Student's' problem when several different population variances are involved, *Biometrika, 34*(1/2), 2835.

[31] Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *arXiv preprint math/0505527*.