A Multi-Layer Quality Control Tool for Weather-Driven Agricultural Decision Support Models

by

Mark Anthony F. Mateo

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada

November 2008

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

A Multi-Layer Quality Control Tool for Weather-Driven Agricultural Decision Support Models

BY

Mark Anthony F. Mateo

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Master of Science

Thesis advisor

Dr. Carson K. Leung

Author

Mark Anthony F. Mateo

A Multi-Layer Quality Control Tool for Weather-Driven Agricultural Decision Support Models

Abstract

Weather plays an important role in agriculture. This calls for reliable weather data, which in turn runs agricultural decision support systems that help farmers make management decisions about their crops. It is well known that "garbage in means garbage out". The presence of anomalies and errors in a weather dataset can gravely compromise the results given by various applications using them. Moreover, many of these applications cannot handle datasets with missing entries and just simply crash. Thus, to have seamless operation and obtain meaningful, accurate results from agricultural decision support systems, it is essential to have a clean and complete set of data. For my M.Sc. thesis, I propose a multi-layer quality control tool for weather data. This multi-purpose quality control tool enhances agricultural decision support systems and eventually improves the farmer's decision-making capability on the management of his crops. Experimental results on real-life datasets show the positive effects of our tool on the quality control of agro-meteorological data.

Acknowledgements

First and foremost, I wish to express my sincerest gratitude and appreciation to my research supervisor, Dr. Carson K. Leung for his encouragement, patience, invaluable support and sound advice on my research project. During my time at the University of Manitoba, he has been my professor, supervisor, mentor, and boss. Without his guidance and encouragement, this thesis could not have progressed.

This work would also not be possible without the assistance of Andrew J. Nadler, an agro-meteorologist from the Manitoba Agriculture, Food and Rural Initiatives (MAFRI), who shared with me his expertise in agro-meteorology. Special acknowledgement goes to Dr. Paul Bullock in the Department of Soil Science at The University of Manitoba for providing the crucial weather datasets used in our experiments. I would also like to take this opportunity to thank the members of my thesis examination committee members, Dr. Abba B. Gumel and Dr. Peter C.J. Graham, and the chair of my thesis defence, Dr. Helen A. Cameron, for their invaluable comments and suggestions on my thesis.

My graduate studies would not have been possible without the direct or indirect financial support of the following organizations: Manitoba Agriculture Food and Rural Initiatives (MAFRI), the Mathematics of Information Technology and Complex Systems (MITACS), Natural Sciences and Engineering Research Council (NSERC), Manitoba Centres of Excellence Fund, the Department of Computer Science and the Department of Statistics.

During my Masters studies, I had the chance to travel both in Canada and abroad specially in places as far as Hong Kong and New Delhi—places which I never ever thought of being able to step my feet into. I wish to acknowledge the generous financial support of my supervisor, Dr. Carson K. Leung, the Department of Computer Science, Faculty of Science, and the Faculty of Graduate Studies for making these conference trips possible.

Living in Winnipeg in the duration of my studies, I am thankful to have met outstanding people here and they have made my graduate studies an enriching life experi-

ence. Special thanks to Alain, Dan, Lloyd, Anna, Naomi, Chrissy, Yui, Olivier, Alf, Khizar, Shinya, Sanae, Taka, Beatriz, Dmitri, and Natalia and other friends whom I might have failed to mention.

I wish to thank my family—my mom, dad and my sister whose never-ending patience, understanding and support during the last four years of graduate school has surely provided a source of strength and inspiration to help me finish my work. Finally, I wish to thank God Almighty in guiding me throughout all the hardships and challenges, in giving me good health and wisdom as I work everyday for the pursuit of scholarship. Thank you very much.

MARK ANTHONY F. MATEO
B.Sc., University of the Philippines - Diliman, Quezon City, Philippines 2004

The University of Manitoba
November 2008

To all those who believed in me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data mining aims to search for implicit, previously unknown and potentially useful patterns and relationships that might be embedded in stored data. Common data mining tasks include association rule mining, sequential pattern mining, clustering, classification, and outlier detection. While most data mining tasks (e.g., association rule mining, sequential pattern mining, clustering, and classification) focus on finding information about the *majority* of data [LNM02, LLN03, LKH05, LK06, LIC08, LB08, LMB08], it is equally important to identify and examine uncommon or rare events occurring in the *minority* of the data [LTB06, CCV07, LKZC07, NHZC07, YSZ07, ZSGL07] as these events are often as interesting as those common ones. As has been said, "a person's noise could be another person's signal", and this "signal" could be an indication of unusual, exceptional, suspicious or erroneous activities. To this end, outlier detection, a data mining task that focuses on finding information in the *minority* of the data, identifies observations appearing to be inconsistent with the remainder of a dataset.

Regardless of the data mining tasks performed with the data, the quality of the mining results depends heavily on the quality of the input data. A good example is the

relationship between weather data and agricultural decision models. In the history of agriculture, weather has played the most influential role in each farmer's decisions regarding the management of his crops. Farmers consult the weather to plan when to seed, irrigate, weed and apply their fertilizers and pesticides. Recognizing the influence of weather in agriculture, scientists have developed various agricultural decision support models such as the potato Disease Severity Value (DSV) [Wal62], the Standard Precipitation Index (SPI) [MDK93], the US National Agricultural Decision Support System (NADSS) [GDH+04], the Palmer Drought Severity Index (PDSI) [WGH04], and the Erosion Productivity Impact Calculator (EPIC) crop growth and yield model [DHJ+04], all of which help farmers decide how best to take advantage of favourable weather conditions for farming and to minimize the adverse effects of unfavourable ones. Weather data are the core driving force behind these models. It is well known that "garbage in means garbage out", meaning an erroneous input may result in an erroneous output. Besides erroneous data, missing and discontinuous entries create incomplete datasets causing some of these agricultural models to fail. In fact, most agricultural decision models require a complete set of data, and they just simply crash when the input dataset is incomplete. Since the accuracy of these agricultural models inevitably relies on the quality of their inputs (i.e., weather data), it is therefore crucial to (i) remove any data deemed erroneous and to (ii) ensure a complete set of inputs prior to data mining. Otherwise, the mining results may be useless or misleading, which in turn may affect decisions made using the application for which the data mining tasks were performed.

In the province of Manitoba, Manitoba Agriculture, Food and Rural Initiatives (MAFRI), the provincial government agency responsible for the well-being of Manitoba food producers, operates a real-time weather monitoring network that collects weather data from the vast agricultural farm lands of the province. With the data from its weather network, MAFRI provides relevant crop reports and agricultural decision support models that rely

on the current and past weather data. Among these models are the potato Disease Severity Value (DSV) model [Wal62] used by farmers as a guide for when to administer fungicides in combating the potato late blight disease, a serious, yield-debilitating disease in potatoes, and the corn Crop Heat (CH) model [GR58] that determines the best corn variety to be cultivated in a given farming region. As these models are dependant on weather observations as their inputs, the more accurate the data, the more reliable is the model output. Hence, data *quality control* (QC) is necessary to scrutinize and identify anomalous observations in the data to make it reliable and error-free.

Considering the widespread reliance of farmers on agricultural decision support models for critical farm management decisions, it is fundamental to have a clean set of weather data for the weather-driven models to use. Having a clean set of data removes doubt from them and ensures confidence in the decision made by farmers based on their weather-driven models. Surprisingly, as we shall discuss in the next chapter, while the literature is *abundant* on outlier detection for general data, it is *limited* on specific outlier detection techniques for the quality assessment of *weather data.* There have been existing methods made in the past for controlling the quality of weather data but most of them are simple, fixed rule-based procedures that do not use outlier detection techniques. In addition, the majority are limited to only single-station quality control. To be more accurate in ensuring the quality of weather data, it would be better to consider not only a single station but also its surrounding stations and use multi-station quality control techniques.

The requirement to remove erroneous observations to obtain a clean and complete set of data inputs and ensure successful data mining leads us to the following questions: (a) How can we develop an automated system for detecting abnormal weather observations using outlier detection methods to control the quality of MAFRI's weather data and enhance its agricultural decision support models? (b) How can we improve ex-

isting data quality control methods to enhance weather-driven decision support models? (c) How can we apply outlier detection techniques on the quality control on weather data and fill-in discarded and missing observations to maintain a complete set of data? (d) How can we integrate incompatible weather data formats into compatible ones for collaborative use? (e) How can we test the effectiveness of our proposed methods? (f) Finally, how can we generalize our tool and extend its applicability from weather data to other diversified application domains?

1.1 Problem Statement and Contributions

In this thesis, we answer the above questions. Specifically, the thesis statament is as follows:

> We develop a quality assurance and control tool to enhance weather-driven decision support models.

The *key contribution* of this thesis is the development of a multi-layer quality control tool that aims to enhance weather-driven agricultural decision support models. To elaborate, our quality control tool is divided into three quality control layers (*internal, temporal and spatial layers*), with each layer specializing in the detection of a particular type of error that might be embedded in the datum. More specifically, to improve on existing single station quality control techniques (provided by the internal layer, we use both *spatial* checks (provided by the spatial layer) and *temporal* checks (provided by the temporal layer) that uses *outlier detection* techniques in the quality control algorithms. *Spatial* checks compare a station's measurement with multiple surrounding stations while *temporal* checks compares a station's present measurement with its previous archival measurement. Both these techniques, namely the spatial and temporal checks, are non-deterministic, meaning the pa-

rameters in which each datum are evaluated are not based on fixed rules but are calculated from the historical data of the stations being considered. Using non-deterministic quality control methods delivers an important improvement from previous methods as these new techniques provide flexibility because users are allowed to dynamically change the assurance level parameters and control the sensitivity and assurance level as desired for detecting outliers. Besides *distance*, we also use *measures of statistical agreement* in calculating the quality control parameter to solve the limitations of previous methods that perform poorly in areas experiencing micro-climates due to the presence of different geographic features.

We emphasize that besides scrutinizing erroneous weather data, our developed quality assurance and control system is capable of filling in data gaps resulting from a discarded erroneous observation or a missing datum and is able to maintain a complete time series dataset. This is specifically helpful for agro-meteorological decision models that can only work when there is a complete time series dataset input. Moreover, we integrate traditional single-station and multi-station techniques for the quality control of weather data. Finally, we discuss how our developed tool can be extended in other application domains besides weather data and provide examples of diversified application domains outside weather data quality control.

Figure 1.1 shows the role and position of our multi-layer quality control tool in the agro-meteorological decision support process.

This thesis is a culmination and extension of my MITACS internship with the Manitoba Agriculture, Food and Rural Initiatives (MAFRI) to improve the quality of its agricultural decision support models for Manitoba farmers. As a result of this internship, some of the ideas presented in this thesis have been published in four refereed papers for international conferences ([LMN07a, LMN07b, ML08a, ML08b]).

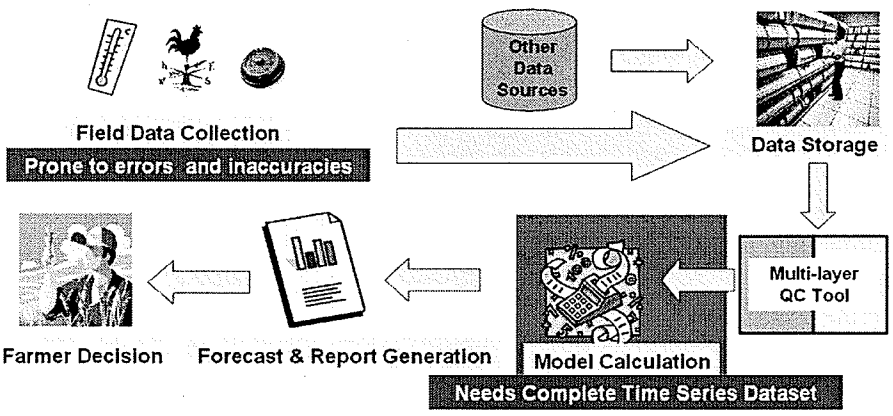


Figure 1.1: Our multi-layer QC tool and its position in the agro-meteorological decision support process.

In this thesis, we focus on detecting outliers in weather data, with particular emphasis on methods involving multi-station and multi-time frame quality control techniques. We include a detailed treatment of multi-station quality control techniques and guidance on replacing missing and outlier data to maintain a complete and consistent time series dataset (Chapter 3). New materials not appearing in previous published works include experimental results to study the ability and performance of the quality control system in (a) flagging erroneous weather observations and (b) accurately calculating an estimate to replace flagged weather data and maintain a complete time series dataset (Chapter 4). In summary, our key contributions are:

1. We develop a quality assurance and control tool that enhances weather-driven agricultural decision support models. Our tool uses multi-station quality control techniques that improve on previous methods by considering surrounding stations in verifying the correctness of a given station's data.

2. Besides checking the quality of data, we develop techniques for filling in data gaps brought about by either a discarded erroneous observation or a missing datum and

thereby provide a complete set of data.

3. We demonstrate that our tool is applicable not only in scrutinizing the quality of weather data but also in other application domains as well. We show this through discussions of several application domains from diversified fields.

1.2 Thesis Outline

Chapter 2 presents a review of related work. First, it provides an introduction to the concepts of data quality control and outlier detection. Next, it traces the evolution of outlier detection from being a strictly statistics-based technique in the past to an important data mining topic at present. In Section 2.2, we discuss the need to have outlier detection techniques that are specifically suited for spatial data (such as weather data) followed by a survey of recently proposed algorithms and an analysis of their applicability for weather data quality control. Finally, we conclude with a review of existing weather data quality control techniques followed by a discussion of their shortcomings that we address in this thesis.

We start describing the architecture of our current work in Chapter 3. Specifically, Chapter 3 describes the data quality control tool we have developed. As the developed tool is multi-layered, we chronologically describe its architecture based on the order of three implementation layers. For each layer, we elaborate on the tests and the algorithms that it uses for data quality control. We also discuss solutions in scenarios that might pose problems during the execution of the data quality control procedures. Finally, we detail the techniques for replacing flagged erroneous data to be able to maintain a complete time series dataset.

Chapter 4 discusses the results of our experiments in evaluating the effectiveness

of our developed QC tool. We performed experiments to evaluate the capability of the tool to flag erroneous weather data and to test the accuracy of the estimated values when replacing flagged observations.

In Chapter 5, we enumerate the challenges we have encountered in the development of our tool. We discuss problems brought about by using data in different formats or national standards as well as the corresponding solutions we have undertaken to solve them.

Finally, we conclude in Chapter 6 with our research findings and provide some ideas for extensions and future work.

Chapter 2

Related Work

In this chapter, we introduce the concept of data quality control and outlier detection techniques for the quality control of data. We also discuss recently proposed outlier detection techniques that are specifically designed for spatial data—the type of data we are working with in this thesis. Next, we proceed with a review of previous quality control techniques and a discussion of their limitations. We end this chapter with an explanation of the key advantages of our quality control tool when compared with existing techniques.

2.1 Background: Data Quality Control

Data quality control [KC04, BBB06, ZC06, CFG+07, LEBQ07] is the process of scrutinizing data to ensure that they are clean and error-free and meet quality requirements such as accuracy, completeness, timeliness, continuity, and consistency [BP85, Red01, HH02, KY05]. In performing data quality control, we do not expect that the data will be 100% error-free. Instead, we demonstrate the quality of the data and make sure that this quality is appropriate for its intended application. To perform data quality control on weather data, we need to scrutinize our dataset and remove any abnormal weather observations.

Outlier detection, the process of identifying observations different from most other observations [DJ04], can be applied to identify abnormal weather data so that they can be removed. Specifically, outlier detection is the main principle behind data quality control in weather data.

In the earlier days, outlier detection was primarily a branch of statistics which relied on model-based approaches in detecting outliers [Haw80, DG93, RL03, Sco03, Yan06]. In the model-based approach, any observation falling outside the model is considered as an outlier. However, over the past decades, outlier detection has attracted substantial attention in the data mining community and has indeed become one of its most important areas. Using data mining principles, outlier detection has been recently applied in diversified application domains ranging from credit card fraud detection [WHA+08] to discovering abnormal highway traffic patterns [SLZ02]. An important application area in which outlier detection has been widely investigated and applied is in intrusion detection in computer networks [EAP+02, LEK+03, LKS05].

2.2 Outlier Detection in Spatial Data

The existence of huge amounts of spatial data has driven the need to have outlier detection algorithms that specifically identify outliers in the spatial context. This is because spatial data are distinct from non-spatial ones since they have additional attributes regarding their spatial information such as location, boundaries and direction. Due to these additional attributes, detecting outliers from spatial data is different from non-spatial ones. The reason is that, to be a spatial outlier, a datum needs to deviate only from its immediate spatial neighbours. In contrast, to be a regular (i.e., non-spatial) outlier, a datum needs to deviate from the majority of data in the entire dataset. Moreover, these additional

attributes increase the dimensionality of the data, making traditional non-spatial outlier techniques ineffective due to the high-dimensionality of the data.

There have been many studies in outlier detection algorithms specifically for spatial data. In particular, Lu and his group [SLZ02, LCK03, LKZC07, CLK08] have focused on improving spatial outlier mining. First, they designed a suite of algorithms [LCK03] that detect spatial outliers using a multi-iterative approach with the median (instead of the more commonly used mean) as the neighbourhood function. This approach made the algorithms more robust and solved the problem of previous spatial outlier algorithms where false spatial outliers were identified while leaving true spatial outliers ignored. Then, to address spatial outlier detection for graph-modelled datasets, Shekhar et al. [SLZ02] proposed an algorithm that uses graphs to represent the topological space in which the spatial data are collected. In this graph-based set-up, a node corresponds to a sensor or a station, and an edge represents the connection between a station and one of its surrounding stations. Moreover, their most recent work [CLK08] improved their original algorithm and added capabilities for detecting multiple attribute outliers (via the Mahalanobis distance-based algorithm). As a major achievement, the algorithm is able to solve the problem of misclassifying normal objects as outliers when their neighbourhoods contain real spatial outliers with very large or very small attribute values.

Apart from graph data sets, algorithms for detecting spatial outliers in trajectory data have also been studied. For instance, Lee et al. [LHL08] proposed a partition-and-detect framework outlier detection algorithm for trajectory datasets. To detect outliers, the outlier detection phase uses a hybrid of distance-based and density-based approaches to give it the ability to detect outlying sub-trajectories from a trajectory database. Their algorithm has been applied and tested in the field of meteorology, and has proven to be effective in the detection of abnormal hurricane trajectories in the Atlantic.

With the recent attention and results that outlier detection is getting, it is important to note that outlier detection is not only a branch of statistics but has evolved over the years to become one of the most important branches of data mining. However, it is interesting to note that, while the scientific literature contains abundant general guidance on outlier detection—which is the core principle of data quality control—the availability of literature concerning the application of these techniques to the quality control of weather data remains scarce. In the remainder of this chapter, I briefly discuss available literature on the quality control of weather data and analyze the weaknesses and limitations of that work, which we address in the course of this M.Sc. thesis work.

2.3 Weather Data Quality Control: Single-Station Techniques

Many weather data quality control studies dealt with straightforward single-station checks, which did not use any outlier detection techniques. For instance, Reek [Ree92] used *static* checks to flag conspicuous errors that discredit meteorological observations and archives. Their quality control procedures are based on fixed, deterministic and non-computational rules. These rules specify that the data should follow a logical pattern with respect to its related attributes. The rules only consider aggregates and summaries of data (such as mean, minimum and maximum) to evaluate if a data observation is valid or not. With these fixed rules, data flagging is straightforward and no computationally complex procedures need to be implemented. While straightforward and easy to implement, their method is limited to analyzing only a *single station* (say, only the station of interest) and fails to consider data observed in neighbouring stations. In contrast, our tool considers both the station of interest and its neighbouring stations (i.e., multiple stations).

On the other hand, Shafer et al. [SFA+00] believed that quality assurance tech-

niques must be applied to ongoing data collection to be able to identify problems before they become serious. In their work, quality assurance techniques are divided into five separate test routines—namely, range, step, persistence, spatial, and like-instrument comparisons. Their self-calibrating persistence test uses the aggregated mean and standard deviation to flag a particular dataset. However, as the persistence test uses aggregate statistics, it unfortunately cannot discern which particular observation within the time period included in the test is an outlier. As a result, to identify which of the observations are outliers, all data values that are used to calculate the aggregated mean and standard deviation need to be flagged and then individually examined. Therefore, if, for example, 1-month data were used to calculate the aggregated quality control parameters, then the manual quality control inspector would have an arduous time determining and isolating which observations within the time period are responsible for the offence. This situation does not occur in our proposed tool as it does the flagging on an individual, but not on a collective basis allowing us to instantly pinpoint which particular daily observation has been determined as an outlier.

2.4 Weather Data Quality Control: Multi-Station Techniques

Most of the techniques that we have discussed so far only deal with single-station quality control techniques. Over the past few years, there have been a few studies that worked on multi-station data quality control. For instance, Kondragunta and Shrestha [KS06] proposed methods to spatially check the consistency of a station with respect to the measurements of its neighbouring stations. Unlike other previous methods, Kondragunta and Shrestha have spatial quality control methods that verify the spatial consistency of a station's measurement with respect to its surrounding stations. Instead of choosing stations

based on distance, a station is included if it falls inside a 1° latitude × 1° longitude inclusion box of the station of interest. The spatial consistency checks are novel as they avoid using the mean and standard deviation (parameters which are easily skewed by outliers), but instead use more robust measures that are not easily affected by outliers such as the median and percentiles. The algorithm also gives users the flexibility to dynamically assign parameters to set the level of quality control assurance desired, but unfortunately falls short of stating guidelines on selecting the threshold values to use. With this, inexperienced users not familiar with the climatology of a particular area may have difficulties in setting an appropriate threshold value. Lastly, Kondragunta and Shrestha's method is limited to *only* checking the spatial consistency for rain gauge data, but *not* for other weather parameters (e.g., temperature, relative humidity) that are equally important in many real-life situations. In contrast, our tool specifies a definite step-by-step guideline on station selection that does not affect the result based on the user's experience or familiarity with the area.

To address the limitations of previously proposed techniques, both Eischeid et al. [EPD+00] and the Norwegian Meteorological Institute (*Det norske meteorologiske institutt*, or DNMI for short) [JFM+02] proposed methods that consider observations on surrounding stations (i.e., methods that allow the spatial QC of weather data). Eischeid et al. proposed the *inverse distance weighing* (IDW) technique, which computes an estimate of the station of interest by considering data observed at surrounding stations and weighing these observations according to the inverse of the distance between these stations and the station of interest. However, the IDW technique relies on the wrong hypothesis that the nearer the neighbouring station is to the station of interest, the more similar are the data observed at these neighbouring stations to those observed at the station of interest (i.e., the better it is as an estimator). Unfortunately, the IDW technique does not account for differences brought about by topographical factors such as elevation making it perform

poorly in mountainous and valley regions.

On the other hand, the method proposed by DNMI allows for spatial quality control of observations. It recommends considering 12 neighbouring stations and performs spatial validation on them. However, it does not provide guidelines on which 12 out of the total number of surrounding stations should be selected. Instead, it gives the prerogative to the meteorologists based on their "experience" and local knowledge of the climatology of the area. Different selections of stations may lead to different results as it is unclear for a non-expert how to determine the "best" selection without exhaustively examining most or all the combinations of the available surrounding stations. This is particularly a problem in areas where there are a dense number of neighbouring stations.

Specific studies where spatial outlier techniques are applied in detecting outliers in weather datasets are very scarce. Recently, Lu et al. [LKZC07] used a spatial outlier detection technique to track regional outliers in a sequence of meteorological image data frames. The algorithm uses the Mexican Hat and Morley transformation techniques to filter out noise and enhance data variation. Their techniques were successful in detecting outliers in spatial satellite image data through the automated identification of hurricanes. However, their algorithm is not applicable to our specific application domain because it only allows *continuous* spatial image data as inputs and does not accept *discrete* spatial data.

Most of the existing published quality control (QC) studies are focused on single-station procedures. There is a lack of work on multiple-station quality control procedures (which would result in more accurate data when combined together with single-station quality control). As well, most of the studies only propose QC procedures and fail to describe what to do when data are discarded or an entry is missing.

2.5 Summary

Our survey of related work reveals that outlier detection, which is the core principle of data quality control, has evolved from being a branch of statistics to an important topic in data mining research. From our literature review, we realized that outlier detection techniques have become specialized over the years. This specialization can be seen on many new outlier detection algorithms that only work with a specific data type (e.g., spatial data). However, while our survey reveals that a considerable amount of work on outlier detection is available for general data, specific techniques designed for the quality control of weather data remain *scarce.*

In addition to outlier detection techniques, we discussed existing data quality control techniques for weather data. The survey revealed that single-station techniques use fixed and not parameter-based criteria, an approach which gives more flexibility and accuracy. We also discussed existing multi-station techniques. However, despite the improvements multi-station techniques bring over single-station techniques, the survey reveals that some multi-station techniques are only confined to a single weather attribute (e.g., [KS06]), are unclear with station selection guidelines (e.g., [JFM+02]), and are only applicable to a particular climate zone (e.g., [EPD+00]). Furthermore, none of the multi-station techniques is able to replace flagged and missing values to maintain a complete dataset.

In this thesis, we address the aforementioned limitations. To complement existing single-station quality control techniques, we use multi-station data quality control procedures. These multi-station QC procedures have clear and more exact guidelines on selecting the stations, which previous methods fail to address. We integrate both the single-station and the multi-station methods to build a tool that remove outliers in weather datasets. Finally, we replace discarded and missing data to maintain a complete set of observations

Table 2.1: Comparing our QC system with previous QC studies.

	Reek et al. [Ree92]	Shafer et al. [SFA+00]	Kondragunta & Shrestha [KS06]	Eischeid et al. [EPD+00]	DNMI [JFM+02]	Our QC System
Single station QC	√	√			√	√
Multi-station QC			√	√	√	√
Multi-attribute QC	√			√	√	√
Dynamic & non-rule based QC		√	√	√	√	√
Fixes erroneous & missing data						√
Systematic station selection criteria			√		√	√

in the weather dataset.

Table 2.1 compares the strengths and limitations of the quality control systems we have discussed with the one we developed in this thesis.

Chapter 3

Our Data Quality Control System

In the previous chapter, we reviewed existing work and analyzed their respective limitations. We also discovered from our review that there was a general lack of outlier detection techniques for weather data. In this chapter, we describe our work. We address these problems by: (i) designing a multi-layer quality control tool for detecting abnormal weather observations; (ii) providing a multi-station QC algorithm that incorporates a systematic criteria for selecting stations and (iii) devising methods for replacing discarded and missing data.

The remainder of this chapter is organized as follows. First, we discuss the overall architecture of our tool. Next, we proceed with an individual discussion of implementation layers of the tool. This will explain in detail all the tests and algorithms used by each implementation layer. We also discuss solutions we chose for situations that posed problems during the execution of the data quality control procedures. Finally, we describe our techniques for calculating estimates to replace erroneous data and maintain a complete time series dataset.

3.1 An Overview of Our Data Quality Control Architecture

In this section, we explain the overall architecture of our data quality control tool. Figure 3.1 presents a schematic diagram of the sequence of quality control procedures.

At the centre of the diagram is a rectangle that represents the main quality control (QC) module. Data (shown in the left portion of the diagram) is fed to the QC module where quality control procedures execute. If a datum does meet the quality requirements, it is then considered clean and ready for use by our agricultural decision support tools. Otherwise, it is flagged and passed to the data filling and estimation module for further processing.

On the left side of the diagram are the data sources. The QC module receives data from diversified sources. For our particular case, we have three sources: (i) Manitoba Agriculture, Food and Rural Initiatives (MAFRI), (ii) Environment Canada and (iii) US data. The requirement of our QC procedures is to use data from surrounding weather stations. The unique geographical proximity of MAFRI weather stations to the US border necessitates the use of data from these three sources. Nevertheless, as our data come from diversified sources, issues arise with respect to compatibility that prevents raw data from being readily used. We further discuss these issues in Chapter 5.

As illustrated in Figure 3.1, the main QC module is divided into three layers. We divided the QC module into layers such that the quality control procedures can be executed in stages, with each stage corresponding to a specific task performed by each layer. The implementation layers are as follows:

- **Internal Layer**: quality control of weather observations at a *single* station and on a *single* time frame.

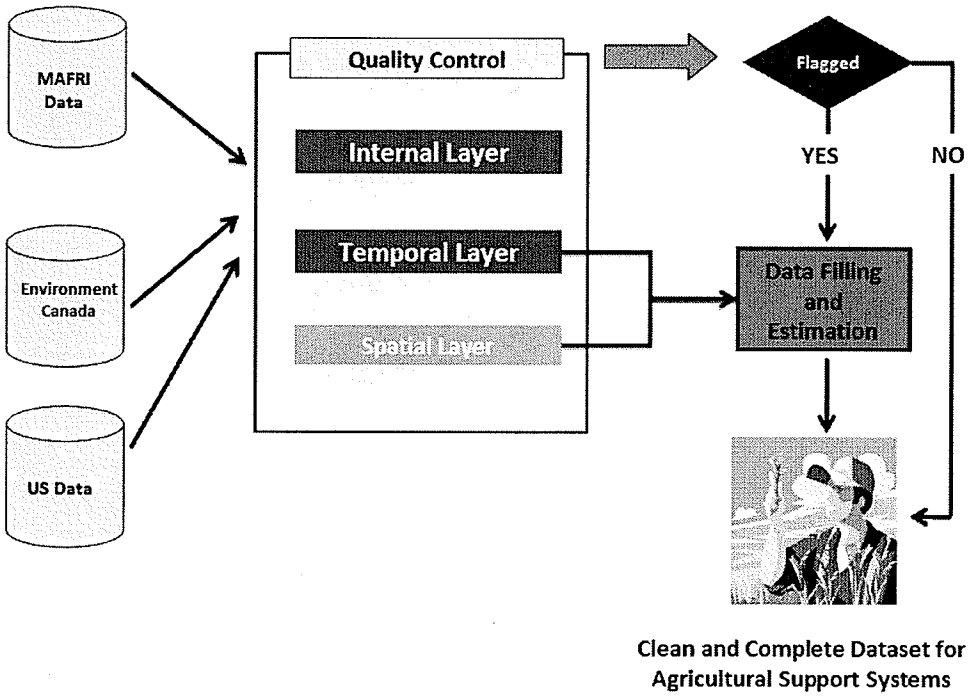


Figure 3.1: Schematic diagram of the data quality control sequence.

Table 3.1: Our prototype's implementation layers and their corresponding functionalities.

Implementation Layer	Functionality
Step 1: Internal Layer	Ensures data lie within reasonable ranges
Step 2: Temporal Layer	Ensures data follows some sequential cycles or patterns
Step 3: Spatial Layer	Checks that data are consistent with their neighbours

- **Temporal Layer**: quality control of weather observations at a single station but on *multiple* time frames.

- **Spatial Layer**: quality control of weather observations at *multiple* stations and on *multiple* time frames.

Table 3.1 summarizes the key functionalities of the three implementation layers.

To be able to pass as "clean", each datum must clear all three QC layers. If a datum passes as "clean", it is ready for use by our agricultural support system. If a datum is flagged and determined to be erroneous, it needs to be discarded. This particular

scenario opens a gap in our dataset which is undesirable, as some agricultural decision support systems do not work with missing data. Hence, to fill in the gaps as a result of a discarded, erroneous or missing datum, we have the *Data Filling and Estimation Module* that calculates estimates for these data. Sections 3.2—3.5 discuss each of the quality control and data estimation modules in greater detail.

3.2 Internal Layer

Observations coming directly from stations are expected to have inaccuracies due to typographical errors and transmission errors, device sensor miscalibration, and other gross forms of errors. The internal layer removes strange data due to these types of errors before subjecting the data to more complicated quality control procedures. In addition, this layer also verifies whether or not the data fall within a valid range as defined by both physical and device measurement extremes and the climatologic zone of the area. To clean and pre-process the data, the internal layer performs the following series of tests: (1) *limit check*, (2) *consistency check*, (3) *excess diurnal range check* and (4) *flat line check*.

Test 1: Limit Checks

Limit checks allow us to identify if the data falls inside a valid and acceptable range as defined by the capacity of the sensors and the maximum (or minimum) physically possible or naturally occurring extremes. During these checks, the tool compares the observed data against extreme values based on the climatologic records for the monthly high and low measurements in a specific area. It flags any observation that does not fall in the following set of interval values:

- Either the observed maximum or minimum temperature values fall outside the record

range of $-60°C$ and $+60°C$.

- Either the observed maximum or minimum relative humidity values fall outside of the possible range of 0% and 100%.
- Either the observed maximum or minimum rainfall values fall outside of the record range of 0 mm and 190 mm.

Values passing the above interval limit check are then compared to the values in the preceding and succeeding time sequences. The tool flags any observation not satisfying the following set of relationships:

- The observed value (e.g., observed temperature) exceeds the highest readings of the previous, the current and the next months.
- The observed value (e.g., observed temperature) falls below the lowest readings of the previous, the current and the next months.

Test 2: Consistency Checks

Consistency checks, the next in the *internal layer's* series of checks, allow us to ascertain that summary values (i.e., average, minimum and maximum values) follow a valid and logical pattern with respect to other summary values for the dataset. The checks verify the relationship among the summary values and ensure that the relationships among them make sense. For example, take temperature as a parameter. Consistency checks ensure that all of the following conditions hold:

- The daily minimum temperature does not exceed the daily mean (average) temperature (i.e., $Temp_{min} \leq Temp_{avg}$).

- The daily maximum temperature does not fall below the daily mean (average) temperature (i.e., $Temp_{avg} \leq Temp_{max}$).

In other words, consistency checks ensure that the following logical condition is satisfied: $Temp_{min} \leq Temp_{avg} \leq Temp_{max}$. Besides temperature, similar verifications are applied to other weather parameters such as rainfall and relative humidity.

Test 3: Excess Diurnal Range Checks

Excess diurnal range checks ensure that the range of consecutively observed values falls under a valid and plausible threshold. For example, observations having daily $Temp_{max} - Temp_{min} \geq 24°C$ will be flagged as it would be representing a very unusual weather phenomenon in Manitoba. Unlike the previous checks where flagged data could be readily discarded without further analysis, flagged observations under this layer need to be subjected to further analysis by a domain expert familiar with the climatology of the area. When a datum is flagged, it is not automatically discarded as two possibilities exist—either the datum is (a) a *truly exceptional* one and not an error (in which case it should be kept) or (b) it is an invalid datum (that should be discarded). In other words, the excess diurnal range checks identify possibly erroneous data with an extraordinarily large daily range between their minimum and maximum readings.

Test 4: Flat Line Checks

Flat line checks are the last in the series of procedures used by the Internal Layer. Our tool examines observed values for their serial continuity and checks if observations remain constant for some consecutive days. Note that it is quite unusual to have a long period of days with the same maximum or minimum temperature. A constant temperature reading for a considerable number of days may be a signal of a malfunctioning temperature sensor.

Of course, there are some exceptions to this. For example, while we flag a sequence of seven or more consecutive constant non-zero rainfall measurements, we do not flag sequences of seven or more consecutive zero precipitation measurements. While it is possible (and not uncommon) to have a long period of dry days, it sounds suspicious (though not impossible) to observe a long period of rainy days, particularly with identical daily precipitation measurements (e.g., daily rainfall of 100 mm for 10 consecutive days).

3.3 Temporal Layer

Once data has been checked for internal errors (e.g., gross errors, relationship inconsistencies with respect to related data, large changes from previous daily values), our tool checks for its consistency with respect to its previous behaviour. In the temporal layer, our tool determines whether the present observation is climatologically consistent with a time series spanning observations in the past. For example, it would be uncanny to have a reading of $-45°C$ in Winnipeg during the month of July as it is in the middle of summer. To flag this possible inconsistency, the tool considers observations from the past, which will manifest a certain degree of consistency brought about by seasonal variation.

In the temporal layer, the tool constructs a time series to evaluate each daily observation. To construct the time series, we need to work with datasets spanning several decades in the past. For each datum x_t obtained on a certain date t, our tool considers data (i) for the *day preceding* t, (ii) for the *day succeeding* t, and (iii) for these three days in all *previous and subsequent years.* For example, suppose we have an archive of daily temperature data spanning from 1950–2008. If date t is 10 February 2007, then my tool considers 9 February 2007 (*day preceding* t), 11 February 2007 (*day succeeding* t) and 9–11 February 1950–2006 (*three days in all previous year*) and 9–11 February 2008 (*three days*

in all subsequent years).

Although we used a yearly cycle in our example above, the temporal layer is applicable to other time resolution cycles (e.g., daily, weekly, monthly quarterly, seasonal or some other cycles). For example, suppose we have an archive of data collected on an hourly basis that follows a weekly cycle. If t is 12:00 noon on Monday in Week 25, we will consider the hourly observations collected at (i) 11:00 am on Monday, (ii) 1:00 pm on Monday, and (iii) 11:00 am, 12:00 noon and 1:00 pm for all the days prior to, and after that week (i.e., Weeks 1–25 and Weeks 26–52). This example shows that, in some applications, we need to include subsequent data (in addition to the past data) when constructing the input time series. The inclusion of subsequent data allows the tool to be applicable to other time resolution cycles.

After obtaining the time series, the temporal layer computes the mean and standard deviation of the time series. In this type of test, we need parameters that are not easily skewed by outliers. Hence, instead of using the traditional mean $\overline{x}$ and standard deviation σ, our tool uses a slightly modified bi-weight mean $\overline{x}_{bw}$ and bi-weight standard deviation σ_{bw} parameters [Lan96] which are *resistant to outlier values.* As a result, they are beneficial to our particular case because pre-quality control datasets may have embedded outlier errors in them that skews the traditional $\bar{x}$ and σ parameters if used. Given a time series of a dataset $x_1, \ldots, x_N$, the parameters $\overline{x}_{bw}$ and σ_{bw} are calculated as follows:

$$\overline{x}_{bw} = med + \frac{\sum_{t=1}^{N} (x_t - med)(1 - w_t^2)^2}{\sum_{t=1}^{N} (1 - w_t^2)^2} \tag{3.1}$$

and

$$\sigma_{bw} = \frac{\sqrt{N \sum_{t=1}^{N} (x_t - med)^2 (1 - w_t^2)^4}}{\left| \sum_{t=1}^{N} (1 - w_t^2)(1 - 5w_t^2) \right|}, \tag{3.2}$$

where (i) med is the median of all x_t and (ii) w_t is the weight from daily observation $t = 1, 2, 3, ..., N$.

To compute w_t, let D_t denote the absolute difference between x_t and med and let $midD$ denote the median among all D_t, that is, $midD = \text{median}(\{D_t \mid D_t = \text{abs}(x_t - med)\})$. Then, $w_t = \min\left\{1, \frac{c\ (x_t - med)}{midD}\right\}$ for some constant c.

Next, the tool calculates a standardized Z-score for each daily observation x_t as follows:

$$\text{Z-score} = \left| \frac{x_t - \overline{x}_{bw}}{\sigma_{bw}} \right|. \tag{3.3}$$

Note that the *Z-score* indicates the number of standard deviations a value is away from the mean. We can select a level of assurance for which to flag our observation as suspicious to attain the desired assurance level. The observations that were flagged are further checked if (i) they represent exceptions (in which case, they are kept) or (ii) they are erroneous (in which case, they are discarded and replaced by our estimates).

When constructing the input time series, we require a minimum of 10 data entries for our time series to be a valid input in the preceeding formulae. This is so as a minimum of 10 data entries is needed for our formulae to be resistant to outliers. For a yearly cycle, this translates to at least 10 years of data for each station. Hence, we construct a time series of data spanning several decades in the past (Equations (3.1) and (3.2)). Although we assume that a complete set of time series data is available, data gaps are unavoidable in almost all climate data. They form a significant obstacle in seamlessly running temporal

checks, which require a complete series. The natural issue to address, therefore, is how to deal with data gaps or an incomplete set of available time series data. The following are the strategies that our tool uses when time series data is incomplete.

Strategy 1: Assuring Data Format Compatibility

Since we are dealing with data spanning several decades in the past, it is certain that MAFRI current weather database will not contain all the datasets needed. MAFRI only started its weather monitoring program quite recently. As such, it would not have records of data spanning (say, the past four decades). Because temporal checks in this stage require observations spanning several decades from the past, we need to access external weather archives, which, because of their age, are stored in *flat file legacy systems*[1] (database standards most applicable at that time but are not readily compatible with MAFRI present RDBMS database standards). Here, our data quality control system has an intermediary tool (Chapter 5), which allows data retrieval and usage by presenting a standardized view of the data both from present-day RDBMS and from legacy flat file database systems.

Strategy 2: Assuring Data Resolution Compatibility

Various data sources give rise to weather data with different temporal resolutions. For example, Manitoba Ag-Weather currently records data of very high time resolutions having an entry every 15 minutes. This is not the case in older databases, where we only have records of, for example, daily highs and lows. If lower resolution data (e.g., daily) is requested and only higher resolution data (e.g., every quarter of an hour) is available, we summarize the raw data to the requested resolution before returning it for further processing.

[1]Legacy systems are those systems built using older computer standards and, because of this, lack compatibility with newer systems. They continue to be used today because they still respond to the tasks required by the organizations using them or because their replacement would entail significant costs and/or downtime.

Strategy 3: Station Name Changes

Name changes are not uncommon (e.g., Hull in Quebec was renamed to Gatineau; the northern city of Churchill in Manitoba was formerly called the Prince of Wales Fort; we saw the Indian cities of Bombay and Madras recently renamed to Mumbai and Chennai respectively). In cases where names of stations get changed, we may have a series TS_1 for a station named Stn_1 prior to date d and another series TS_2 for a station named Stn_2 after d (where Stn_1 was renamed to Stn_2 on date d). Here, TS_1 and TS_2 are disjointed. To solve this problem, we, together with the help of domain experts who are familiar with station name changes, combine the two related time series together into one.

Strategy 4: Closing of Old Stations and Opening of New Stations

Data might be unavailable for certain periods of time. Due to various reasons (e.g., budget cuts, changes in agricultural activities, land development), some new stations open and some existing stations may close down. For example, when a station Stn_3 is opened on date d, we only have data for this station on or after date d. We do not have data for station Stn_3 before date d. Similarly, when a station Stn_4 is closed effective on date d', we only have the data for Stn_4 before date d' but not on or after d'. To solve this problem, we use data from neighbouring stations.

Strategy 5: Log of Data Absences

For cases where data are genuinely absent for the desired period and queries were not successful in returning any data, the intermediary tool (Chapter 5) records this absence and creates a log of the actual presence or absence of data in the different databases. In the temporal layer, the intermediary tool gradually learns from this log and allows more seamless queries in the future, resulting in quicker responses by skipping the complete execution of

a query that is known not to return any data.

3.4 Spatial Layer

The spatial layer is the last of the three implementation layers in our quality assurance and control tool. The previous two layers only consider a single station of interest and do not examine whether observations are considerably different from the data values in neighbouring stations. We adapted the spatial regression test [HGS+05] as the basis for the tests in the spatial layer. The principle behind this spatial regression test is to calculate an estimate value of the station of interest using the observations of neighbouring stations and to then use this estimate to construct a confidence interval to decide whether the observation at the station of interest deviates significantly from the calculated estimate.

The first step in the spatial layer test is to determine which of the surrounding stations should be included in the test. In other words, how do we select which surrounding stations should be considered, given a station of interest? We can either define a radius limit or a co-ordinate inclusion box wherein stations falling inside the radius or inside the inclusion box will be considered. However, in the first approach, with the number of weather stations available, the distance of one station to all its surrounding station is not usually available. Usually, only the latitude and longitude co-ordinates are provided for each station. Therefore, in our implementation, we used the radius approach and calculated the distance of a station to its neighbours using the respective co-ordinates of the station. In order to get statistically significant results, in our implementation, we used a minimum of 12 stations falling inside a 100-km radius in choosing the preliminary list of stations.

After identifying the neighbouring stations, we need to select which of these 12 stations should be included in the parameter calculation. To do so, the tool determines

the agreement of each neighbouring station with the station of interest. The more similar the observation at a neighbour with that at the station of interest, the better estimator we will have. The spatial layer calculates the correlation coefficient R between the daily observations of the station of interest and each of the 12 candidate stations of the weather attribute being analyzed (say maximum temperature). It then eliminates those stations that are not significant at the 95% confidence level and selects five stations with the highest correlation coefficient R.

In this thesis, we did not use the distance as a measure of agreement to calculate the QC parameter. Instead, we considered the strength of the relationship as quantified by the root mean square error (RMSE) between the station of interest and each neighbouring station. For each of the surrounding stations, we formed a regression-based estimate $\widehat{x}_i$ at date j:

$$\widehat{x}_{ij} = a_i + b_i \, y_{ij}, \tag{3.4}$$

where y_{ij} is the observation of station i at date j, and a_i and b_i are the respective intercept and slope regression coefficients for station i. Since we are using five surrounding stations, we have five regression equations. Data used in calculating each of these regression equations are used for calculating the root mean square error (RMSE) s_i for each of the five stations:

$$s_i = \sqrt{\frac{\sum_{j=1}^{N} (x_{ij} - \widehat{x}_{ij})^2}{N}}, \tag{3.5}$$

where x_{ij} is the value from the station of interest i at a certain date j and $\widehat{x}_{ij}$ is the estimate obtained from station i at date $j = 1, 2, 3, ..., N$. In other words, we sum the difference between the actual and the estimated values for each of the five neighbouring

stations. Next, each RMSE s_i is used to calculate a single weighted estimate w_j:

$$w_j = \frac{\sum_{i=1}^{m} \frac{y_{ij}}{{s_i}^2}}{\sum_{i=1}^{m} \frac{1}{{s_i}^2}}, \tag{3.6}$$

where m is the total number of station-pair regression equations (i.e., station of interest and neighbouring stations, which in our case is equal to five), and y_{ij} is the daily observation at neighbouring station i and date j. The standard error of estimate E is calculated as:

$$E = \sqrt{\frac{m}{\sum_{i=1}^{m} \frac{1}{{s_i}^2}}}, \tag{3.7}$$

where m is the total number of station-pairs as in Equation (3.6).

We note a very important feature of the spatial regression test from the above equations. In calculating the weighted estimate, the spatial regression test does not assign the largest weight to the nearest neighbour. Instead, it looks into each station's statistical agreement with the station of interest. The strength of this statistical agreement is measured by the *root mean square error* s_i values between the station of interest and each surrounding station. Thus, the weighted estimate w gives heavier weight to those stations that have a smaller standard error of estimate, irrespective of the station's distance from the station of interest. The smaller the standard error of estimate between the station of interest and a neighbouring station, the nearer its behaviour to the target station and the heavier weight or influence is given to it in calculating the weighted estimate. This means that the spatial regression test does not use the common assumption that "the nearest neighbouring station best copies the observation of the station of interest (because it is near to the station of interest) and that this is the best estimator". In effect, it is possible to have a situation

where the station that is the farthest from the station of interest is given the heaviest weight in calculating the weighted estimate because it has the best statistical agreement with the station of interest. This is an important feature of the spatial regression test as it compensates for regions manifesting microclimates (i.e., regions observing consistent weather observational variability in just a small geographic space). The large observational variability can be brought about by factors such as topographical features (e.g., mountains, valleys) bringing remarkable elevation differences to nearby stations. For example, we expect consistent variation between the temperature readings of Kananaskis and Morley towns in Alberta. Kananaskis is located along a mountain ridge with an elevation of 1,350 metres, while the town of Morley is just on the edge of the Rockies but still lies on the plains. Kananaskis and Morley are just a few kilometres apart but, despite their close distance, there is a significant consistent difference in the temperature readings between them due to the large difference in elevation. In this case, temperature readings in Kananaskis would agree more with the readings in, say, Banff (a nearby town) than Morley, even though the distance separating them is farther than the distance to Morley. Banff (a ski resort 1,463 metres in elevation) shares the same climatologic region with Kananaskis because of their similar elevation. Since observations in Kananaskis statistically agree more with Banff than the observations in Morley, Banff will have a smaller standard error of estimate in its regression equation and will be given heavier weight by the spatial regression test in calculating the weighted estimate. This will not be the case in other distance-based spatial tests particularly the Inverse Distance Weighing (IDW) technique.

After obtaining the weighted estimate w_j and standard error of estimate E, the spatial layer calculates the confidence interval $[w_j - FE, w_j + FE] \in \mathbb{R}$:

$$w_j - FE \leq x_j \leq w_j + FE, \tag{3.8}$$

where F is a user-specified constant that sets the level of assurance of our quality control procedure.

Each observation is evaluated using the above confidence interval. If the relationship in the confidence interval holds, then the corresponding observation passes the test. Adjusting F allows us to specify the level of assurance desired for the test. Increasing the value of F decreases the number of invalid extreme values that are flagged as we have a larger interval. On the other hand, decreasing F increases the number of valid values that are flagged as we have a "tighter" interval.

3.5 Data Filling and Estimation Module: Replacing Erroneous Data and Maintaining a Complete Time Series Dataset

After the implementation of each quality control layer, the next step is to integrate the layers together and run the quality control tests to identify outliers. We execute the QC procedures in sequence, beginning with the *internal layer*, then *temporal* and finally the *spatial layer*. However, remember that some algorithms, particularly temporal checks in the temporal layer, require a complete set of archival time series data that spans the past few decades to execute the algorithm. Moreover, MAFRI current weather database, our main data source, has just been recently established. Hence, it does not provide all the datasets needed. We therefore need to access external data sources such as Environment Canada's data archive. For this particular case, we used the Canadian National Climate Data and Information Archive from Environment Canada that contains Canadian weather observations since the late 1800s to satisfy the data requirement in generating a complete time series archive of the temporal layer. If any QC procedure identifies an observation as

Table 3.2: A sample output for station SPRUCETCLP (in Spruce Siding, Manitoba) for the Temporal and Spatial Layers.

(a) Output for Temporal Layer

Date	Original	TEst	TDiff	Z-score	TFlag
01 Jan 2000	−2.8°C	−11.81°C	9.01°C	1.17	0
02 Jan 2000	−10.6°C	−12.93°C	2.33°C	0.32	0
03 Jan 2000	−13.9°C	−14.49°C	0.59°C	0.09	0
04 Jan 2000	−20.0°C	−15.29°C	−4.71°C	0.71	0
05 Jan 2000	−16.1°C	−15.74°C	−0.36°C	0.05	0
06 Jan 2000	−15.0°C	−15.48°C	0.48°C	0.07	0
07 Jan 2000	−11.1°C	−15.91°C	4.81°C	0.60	0
08 Jan 2000	−15.0°C	−16.24°C	1.24°C	0.15	0
09 Jan 2000	−13.9°C	−16.43°C	2.53°C	0.29	0
10 Jan 2000	−21.1°C	−16.64°C	−4.46°C	0.57	0
11 Jan 2000	−27.8°C	−16.02°C	−11.78°C	1.63	0
12 Jan 2000	−17.8°C	−15.15°C	−2.65°C	0.34	0
13 Jan 2000	−16.1°C	−14.83°C	−1.27°C	0.16	0
14 Jan 2000	−16.1°C	−14.98°C	−1.12°C	0.13	0
15 Jan 2000	−22.2°C	−15.58°C	−6.62°C	0.78	0
⋮	⋮	⋮	⋮	⋮	⋮

(b) Output for Spatial Layer

Original	SEst	SDiff	SFlag
−2.8°C	−8.87°C	6.07°C	1
−10.6°C	−9.43°C	−1.17°C	0
−13.9°C	−13.94°C	0.04°C	0
−20.0°C	−19.17°C	−0.83°C	0
−16.1°C	−13.82°C	−2.28°C	0
−15.0°C	−14.09°C	−0.91°C	0
−11.1°C	−13.42°C	2.32°C	0
−15.0°C	−14.38°C	−0.62°C	0
−13.9°C	−13.72°C	−0.18°C	0
−21.1°C	−21.68°C	0.58°C	0
−27.8°C	−24.71°C	−3.09°C	0
−17.8°C	−19.73°C	1.93°C	0
−16.1°C	−15.52°C	−0.58°C	0
−16.1°C	−19.96°C	3.86°C	0
−22.2°C	−20.80°C	−1.40°C	0
⋮	⋮	⋮	⋮

an outlier, it is then replaced with an estimate datum generated using the procedures from either the *temporal* or the *spatial* layer. After running the quality control procedure, each station will have an output table similar to Table 3.2.

Consider the columns for the Temporal Layer in Table 3.2(a). The first column, `Date`, records the date of each daily observation. The next column, `Original`, records the original values. Column `TEst` records the temporal estimate $\overline{x}_{bw}$, which is the bi-weight mean parameter (Equation (3.1)) from the temporal layer. Values in `TDiff` are the differences between the temperature data in columns `Original` and `TEst` (i.e., `TDiff` = `Orig` − `TEst`). Column `Z-score` records the Z-score calculated with respect to the original and estimated temporal values. Finally, the column `TFlag` indicates whether or not an observation is flagged as an abnormal one. The column assumes a value of 1 when a datum is flagged, 0 if not flagged, and −1 when the datum is missing. For the example in Table 3.2(a), the Z-score values of all observations did not exceed 3.0 and thus, these observations were considered normal and not flagged. The selection of the Z-score flagging limit of 3.0 ensures that we cover as much data as possble, as 99.73% of values are within

three standard deviations from the mean [Moo04].

Similarly, a sample output for the Spatial Layer is shown in Table 3.2(b). This table is very similar to Table 3.2(a). In the illustration, we omit column `Date` and show column `Original` for ease of readability. Column `SEst` records the spatial estimate w_j, which is the weighted estimate parameter (Equation (3.6)) calculated in the spatial layer. Values in `SDiff` are the differences between the data in columns `Original` and `SEst` (i.e., `SDiff` = `Original` − `SEst`). Column `SFlag` shows whether or not an observation is flagged. For the data in Table 3.2(b), the first weather datum observed was flagged because it fell outside our calculated the confidence interval, whereas the remaining 14 observations were considered normal.

So far, we have discussed how to identify erroneous data. Next, let us discuss how to replace these erroreous data. Specifically, if a data row with `TFlag` = 1 (i.e., the datum is flagged) and is verified to be erroneous, we replace such an erroneous value with the corresponding estimate value `TEst` of such data row. Similarly, if a data row with `SFlag` = 1 (i.e., the datum is flagged) and is verified to be erroneous, we replace such an erroneous value with the estimate value `SEst` of such data row. If a data row with `SFlag` = −1 (i.e., the datum is missing), we fill the gap with the estimate value `SEst` computed based on the values observed at neighbouring stations.

In the next chapter, we will examine which estimation method provides a more accurate estimate. The results will tell us which among the two estimation method (i.e., temporal or spatial) will be more preferable to use in replacing missing data.

3.6 Summary

In this chapter, we described the architecture of the multi-layer quality control tool that we have developed. We also discussed each implementation layer in the quality control module—namely, the internal, temporal, and spatial layers as well as the data filling and estimation module.

The internal layer makes sure that data lie within reasonable ranges. It consists of four tests that each datum needs to undergo—*limit check, consistency check, excess diurnal range check* and lastly, the *flat line check.* The temporal layer accepts a range of data that will be used to ensure that the data follows some sequential cycle or patterns. The spatial layer, on the other hand, makes sure that data are consistent with their neighbours. Finally, by using the temporal and spatial layer parameters, the data filling and estimation module calculates an estimate, which can be used for replacing those flagged erroneous observations and filling those missing observations. As such, we could maintain a complete time series dataset.

In the next chapter, we will empirically evaluate the performance of our developed system in the quality control of agricultural weather data. To be more specific, we will evaluate its effectiveness in flagging and detecting erroneous datasets. We will also compare the calculated estimates with the actual measurements to have an idea of the capacity of our tool in calculating accurate estimates to maintain a complete time series set of data.

Chapter 4

Experimental Evaluation

As one of its main features, our quality control and assurance tool replaces erroneous and missing observations to maintain a complete time series dataset. To do so, we used *temporal–* and *spatial–based* methods as estimators. In this chapter, we evaluate the performance and effectiveness of our tool. We performed experiments on data from stations in the provinces of Alberta, Saskatchewan, and Manitoba. We obtained the datasets, spanning 30 years of daily maximum and minimum temperature observations, from Manitoba Agriculture, Food and Rural Initiatives (MAFRI), Environment Canada, and some US stations adjacent to the Manitoba–US border. We subjected the data to quality control using our multi-layered quality control tool with experiments testing the accuracy of our tool in (a) flagging erroneous observations and (b) calculating estimates of the actual weather data.

4.1 Experimental Set-up

We implemented the data quality control procedures using T-SQL, a specialized scripting language in the MS SQL suite that provides SQL query and logic control. Experiments were conducted on an AMD Athlon machine with a 2.01GHz processor, 4.0GB

RAM and 500GB hard drive.

We classified the experiments into two types based on the performance criteria being evaluated. We performed experiments that measured the following:

a. **Random Seeding Experiment**—the QC system's capability in flagging erroneous data (Section 4.2).

b. **Accuracy of Estimates**—the accuracy of the calculated estimates of the actual observation that can be used to replace flagged or missing values (Sections 4.3 and 4.4).

In particular, we have done the following:

1. We compared the percentage of errors identified with the total number of seeded errors when the dataset is randomly seeded with artificial errors.

2. We examined the tool's accuracy in detecting seeded errors by measuring *precision, recall* and *F-score* (defined shortly) while adjusting the quality control sensitivity levels.

3. We examined the accuracy of estimates through the use of the *mean absolute error* and *root mean square error* tests.

4. We evaluated the goodness of fit of the estimates through the use of the *Pearson correlation coefficient* (R^2) and the *Nash-Sutcliffe Coefficient of Efficiency* (E) tests.

We performed each experiment twice, one for the data field $Temp_{max}$ and another for $Temp_{min}$. We have also performed experiments on other data fields (e.g., humidity, rainfull). To avoid distraction, we only show the results on temperature.

4.2 Experiment Set 1: Random Seeding Experiments

Recall that in the spatial quality control technique (Section 3.4), we constructed a confidence interval (Equation(3.8)) where we evaluated individual daily observations. For each of these evaluations, we decide to either (a) accept the datum as true and valid or (b) flag and label the datum as an outlier. If the datum is valid and is accepted as "valid" or if the datum is invalid and rejected as "invalid", then our tool is working properly. If, on the other hand, a valid datum is "rejected", it is a false negative. If an invalid datum is "accepted", it is a false positive. To have an effective data quality control tool, we want to maintain a balance wherein we reduce the number of "accepted" invalid data while minimizing the number of "rejected"' valid data.

For this experiment, we seeded a 30-year period (1971–2001) of available weather data from MAFRI and Environment Canada with errors of random magnitude. We selected station `STEINBACH` (in Steinbach, Manitoba) and its five adjacent stations (as determined by the algorithm) for the seeding experiment. In the seeding process, we selected both the dates and error magnitude randomly. The new "erroneous" observations are calculated as follows:

$$x_{error_d} = \bar{x}_d + \sigma_{month} * r_i, \tag{4.1}$$

where x_{error_d} is the new "erroneous" observation at random date d, $\bar{x}_d$ is the monthly mean observation in which random date d belongs, σ_{month} is the monthly standard deviation in which random date d belongs, and r_i is a random number with a range of ± 4.0. The selection of 4.0 ensures that our experiment included cases that are close to the extremes of our dataset. We stored and indexed the dates on which the errors were introduced such that we would be able to determine if a particular seeded error has been successfully detected in our experiments.

Evaluation 4.1 (Precision, Recall & F-measure Tests) To evaluate the accuracy of detecting errors, we performed quality control procedures on our seeded dataset and used *precision*, *recall* and *F-measure* metrics in gauging the performance of our quality control tool. *Precision* P is given by:

$$P = \frac{tp}{tp + fp} \tag{4.2}$$

where tp is the number of successfully identified seeded errors (i.e., true positives) and fp is the number of identified false errors (i.e., false positives or values incorrectly identified as errors).

Recall R is given by:

$$R = \frac{tp}{tp + fn} \tag{4.3}$$

where tp is the number of successfully identified seeded errors (i.e., true positives) and fn is the number of seeded errors that were not successfully identified (i.e., false negatives).

We then calculate the *F-measure*, a popular measure for accuracy that summarizes the *Precision* and *Recall* values:

$$F\text{-}measure = \frac{2PR}{P + R} \tag{4.4}$$

The nearer the value of *F-measure* to 1, the more accurate our QC tool is in detecting errors.

In our tests, we adjusted the value of the assurance level parameter F (Equation (3.8)) from 0.0 to 6.0, with increments of 0.5. The selection of our F from 0.0 to 6.0 allows us to have coverage of virtually the entire range of values in our dataset, as an assurance level parameter value of 6.0 covers 99.999% of our data [Moo04]. Recall that the assurance level parameter F is a user-specified constant that can be adjusted to set the

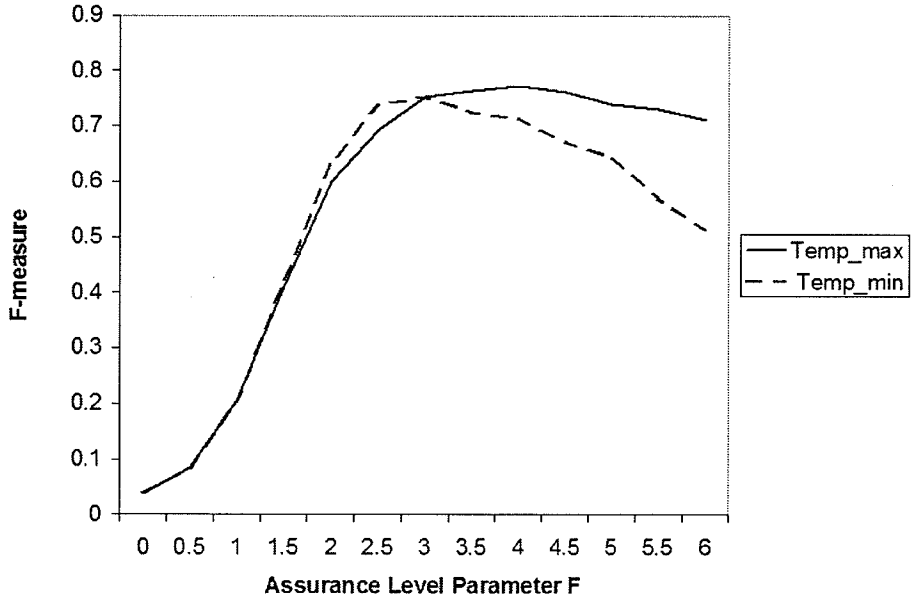


Figure 4.1: *F-measure* coefficient with respect to varying values of assurance level parameter F (Evaluation 4.1).

level of QC assurance and sensitivity. This method of testing against a range of F values makes it possible for us to learn which F value would give the most optimal seeded error detection result (i.e., reduce the number of "accepted" invalid data while minimizing the number of "rejected" valid data).

Figure 4.1 shows the results of our tests. From the graph, our quality control tool performed very well—reaching *F-measure* values as high as 0.75 for both $Temp_{max}$ and $Temp_{min}$. The highest F-measure values occur when the assurance level parameter F is set between 2.5–3.0 for both $Temp_{max}$ and $Temp_{min}$. This is the case where we optimize our result, that is we minimize both the number of "rejected" valid data and the number of "accepted" invalid data.

Evaluation 4.2 (Proportion of Flagged Seeded Errors) Knowing the most optimal value of F, we then proceeded to seed datasets for two other stations namely stations

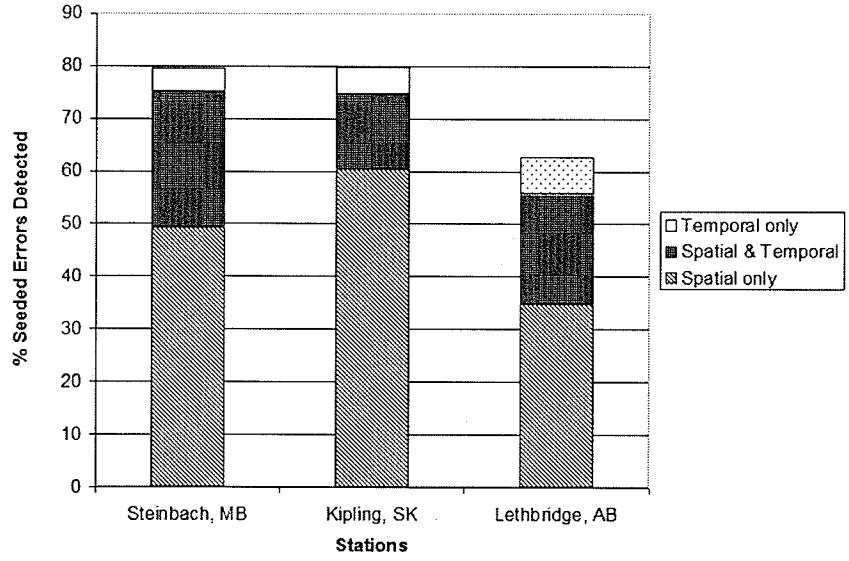
`Kipling` (in Kipling, SK)and `Lethbridge` (in Lethbridge, AB). We subjected the new seeded datasets to quality control using the assurance parameter $F = 3.0$ and recorded the proportion of the seeded errors detected.

It is important to note that we selected station `Lethbridge` to evaluate the performance of the quality control tool in mountainous climate zones, a problematic climate zone in weather data quality control where past QC methods performed poorly (Section 2.4).

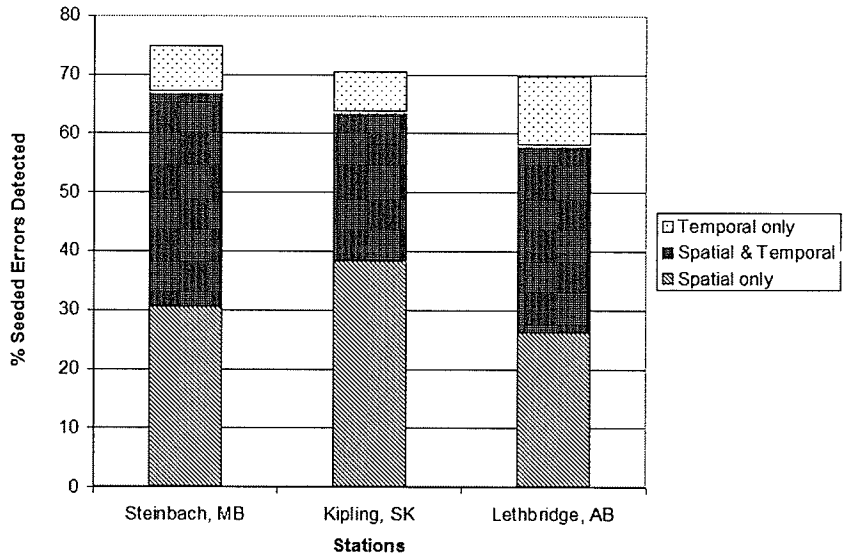
Figure 4.2 shows a bar plot of the result of our seeding experiments for the three stations. For $Temp_{max}$ (Figure 4.2(a)), almost 80 % of seeded errors were successfully detected for the two prairie stations (`Steinbach` and `Kipling`), while the rate for the mountain station was lower, with approximately a 62% detection rate. In general, the majority of the errors were detected by the spatial layer although there was also a remarkable portion of overlaps in detection between the temporal and spatial layers.

For $Temp_{min}$, the same results can be observed with all the stations ranging in the 70–75% detection range. The detection rate for the mountainous station `Lethbridge` was slightly better for $Temp_{min}$ as it registered a 70% detection rate. Compared to the previous graph, we observed that the majority of the detection was an overlap between the temporal and spatial layer.

Those seeded errors that were not detected are those that received very small error magnitudes during the random seeding process. These errors could have been detected with a higher value of the assurance level parameter F. However, as explained earlier, increasing F results in the corresponding increase in false positive detection rate, lowering precision and thereby lowering our accuracy level.



(a) $Temp_{max}$



(b) $Temp_{min}$

Figure 4.2: Bar Plot of the Percentage of Seeded Errors Successfully Detected (Evaluation 4.2).

4.3 Experiment Set 2: Accuracy of Estimates–MAE & RMSE Tests

We used two parameters namely the *mean absolute error* (MAE) and the *root mean square error* (RMSE) to evaluate the accuracy of the spatial and temporal estimation methods. The *mean absolute error* is calculated by the following formula:

$$\text{MAE} = \frac{\sum_{j=1}^{N} |x_j - \widehat{x}_j|}{N} \tag{4.5}$$

and the *root mean square error* is given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} (x_j - \widehat{x}_j)^2}{N}}, \tag{4.6}$$

where x_j is the original value and $\widehat{x}_j$ is the corresponding estimated predicted value (which is the weighted estimate w_j in Equation (3.6)) for data from dates $j = 1, 2, 3, ..., N$.

In the experiments, we calculated the MAE and RMSE for each station for the entire 30-year period and took the average of the MAEs and RMSEs to evaluate the accuracy of the estimates. The lower the MAE and RMSE values, the better our estimates are. We have included an observation in the calculation of MAE and RMSE when it passes its respective quality control layer check (i.e., temporal check or spatial check). To summarize the values for all stations during the entire 30-year period, we averaged the calculated MAEs and RMSEs. The results are presented in Table 4.1.

Evaluation 4.3 (Temporal estimates tabular results) For *temporal QC*, we evaluated more than 190 stations over the 30-year period. To obtain a summarized picture of the accu-

Table 4.1: Average mean absolute error (MAE) and root mean square error (RMSE) for a 30-year (1971–2001) dataset (Evaluations 4.3 and 4.4).

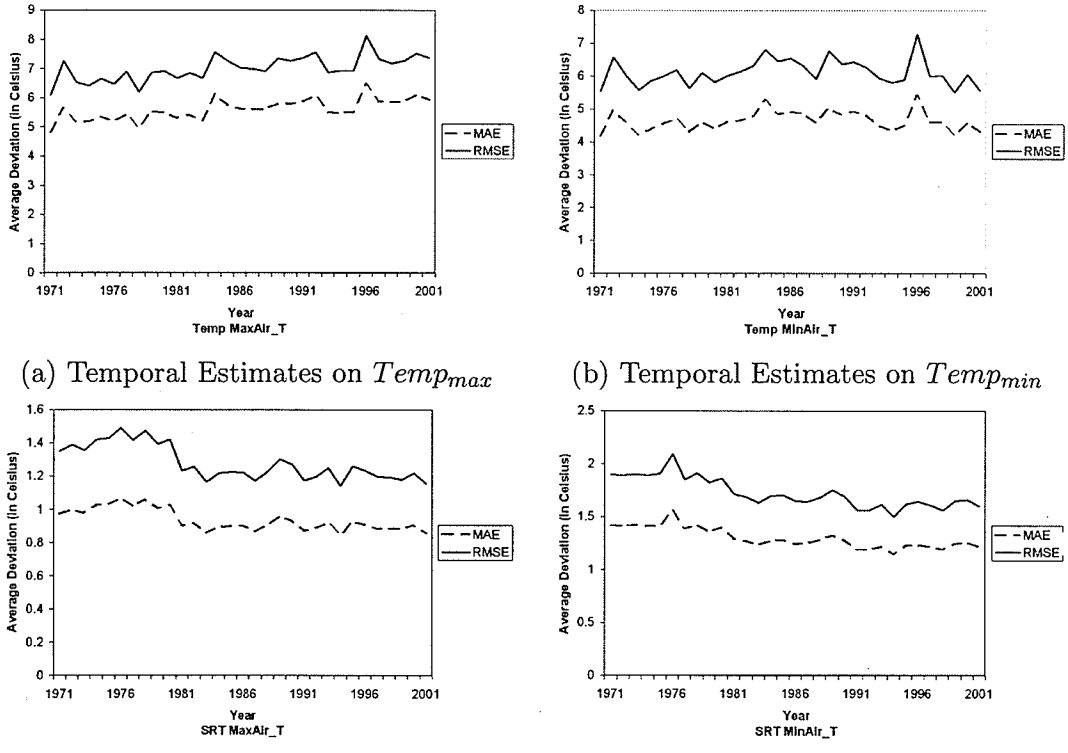
	Temporal Method		**Spatial Method**	
	$Temp_{max}$	$Temp_{min}$	$Temp_{max}$	$Temp_{min}$
No. of Stns	199	190	192	187
Avg MAE	5.46°C	4.63°C	0.98°C	1.35°C
Max MAE	8.68°C	5.41°C	2.41°C	2.66°C
Min MAE	4.53°C	3.70°C	0.50°C	0.83°C
Avg RMSE	6.89°C	6.13°C	1.36°C	1.81°C
Max RMSE	10.47°C	7.61°C	3.39°C	3.56°C
Min RMSE	5.85°C	5.04°C	0.70°C	1.13°C

racy of our estimates, we averaged the MAE and RMSE of all the stations over the available dataset time period. Results showed satisfactory values for both the average MAE (5.46°C) and the average RMSE (6.89°C) for the *maximum* temperature. The same comments apply to the *minimum* temperature with an average RMSE of 6.13°C.

Evaluation 4.4 (Spatial estimates tabular results) Similarly, for *spatial* QC, we evaluated more than 180 stations with their MAEs and RMSEs averaged. Compared to temporal estimates, spatial estimates had lesser stations (192 vs. 199 for $Temp_{max}$ and 187 vs. 190 for $Temp_{min}$). This was because some stations are located in isolated regions with no immediate neighbouring stations available. Recall that in the spatial layer, we need to have a minimum of five neighbouring stations to calculate a spatial estimate. Thus, in situations where a station is located in an isolated area with insufficient immediate neighbouring stations, a spatial estimate cannot be possibly calculated.

As shown in Table 4.1, we obtained very impressive values for both the maximum and minimum temperatures with the average MAE being only 0.98°C and the average RMSE being 1.36°C for the *maximum* temperature. The average MAE and the average RMSE were slightly higher for the *minimum* temperature, registering at 1.35°C and 1.81°C, respectively.

Evaluation 4.5 (Temporal MAE and RMSE time plots) To identify any trend in the accuracy of our estimates over the years, we plotted the yearly average MAEs and RMSEs for the *temporal* and *spatial* estimation methods for both the *maximum* and *minimum* temperatures (see Figure 4.3). For both estimation methods, the trend of the RMSE followed the trend of the MAE, as expected (whenever there was an increase/decrease in RMSE values, there was also a corresponding increase/decrease in MAE). The MAEs for the *temporal* estimators for the *maximum* temperature were slightly high in value, ranging from as low as 4.8°C to as high as 6.2°C (see Figure 4.3(a)). Most of the values fluctuated around the interval [5.0°C, 6.0°C]. The RMSE followed the same trend with slightly higher values. On the other hand, the MAE values for the *minimum* temperature ranged from



(a) Temporal Estimates on $Temp_{max}$

(b) Temporal Estimates on $Temp_{min}$

(c) Spatial Estimates on $Temp_{max}$

(d) Spatial Estimates on $Temp_{min}$

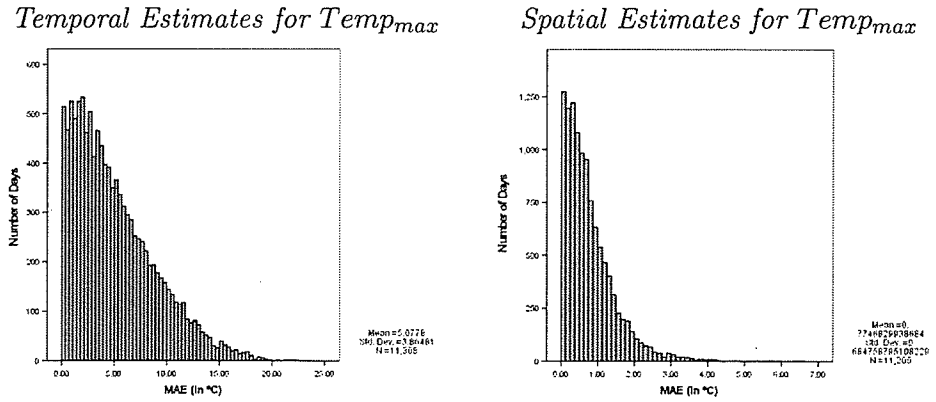
Figure 4.3: MAE and RMSE plots for temporal and spatial estimates (Evaluations 4.5 and 4.6).

4.0°C to 5.5°C with the majority of the values clustered around the interval [4.0°C, 5.0°C] (see Figure 4.3(b)). Similar to the graph for the maximum temperature, the RMSE followed the same trend with slightly higher values.

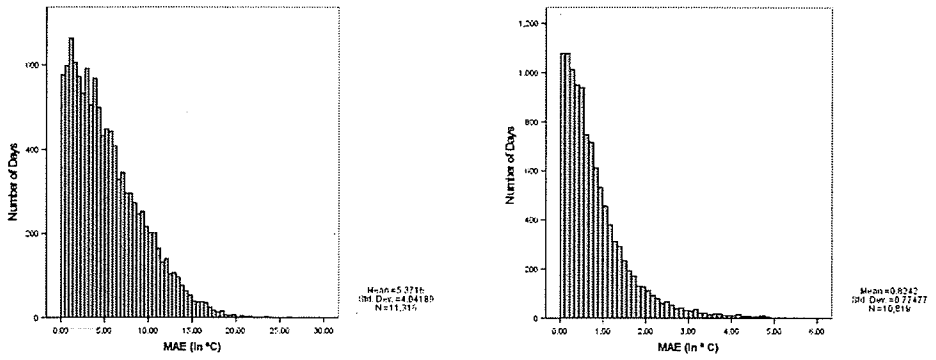
Evaluation 4.6 (Spatial MAE and RMSE time plots) The MAEs for the *spatial* estimators for the *maximum* temperature (see Figure 4.3(c)) hovered slightly above 1.0°C from 1976–1981, and then stabilized to [0.9°C, 1.0°C] after 1981. On the other hand, the MAE values for the *minimum* temperature (see Figure 4.3(d)) hovered around the upper range of 1.5°C to 1.6°C during the mid-1970s to the late 1970s before stabilizing to 1.0°C or below afterwards. MAE values of 1.0°C and below are exceedingly low and represents very impressive estimation accuracy by the spatial estimation method. This trend was also followed by the RMSE with a slightly higher set of values.

4.3.1 Temporal and Spatial Histograms and Cumulative Bar Plots

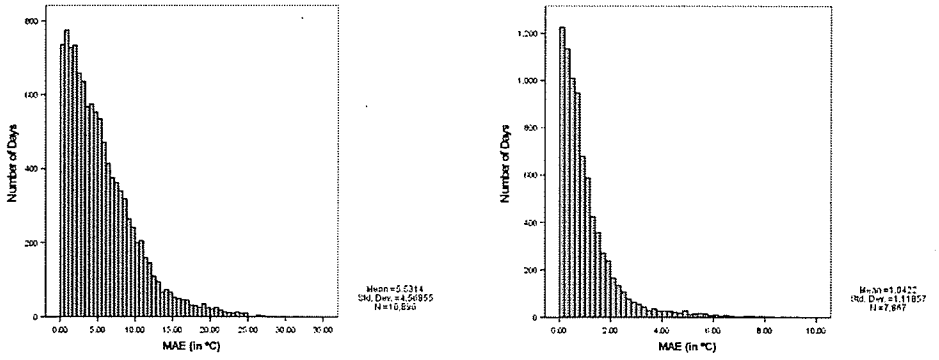
To have a more in-depth analysis of the performance of our estimates, we plotted histograms (see Figures 4.4 and 4.6) and cumulative bar plots (see Figures 4.5 and 4.7) of the mean absolute errors for the entire 30-year period of our dataset (approximately 10,000 data points). In the plots, we selected three representative stations, one each from Manitoba (`Steinbach`), Saskatchewan (`Kipling`), and Alberta (`Lethbridge`). We selected these station such that we represent both the plains (Steinbach, MB and Kipling, SK) and mountainous regions (Lethbridge, AB). This selection enables us to determine the performance of our estimators with respect to stations belonging to different climatic regions. Stations `Steinbach` and `Kipling` are located in typical prairie agricultural lands while station `Lethbridge`, which has an elevation of 929 metres, lies at the edge of the Rocky Mountain range.



(a) Steinbach, MB on $Temp_{max}$



(b) Kipling, SK on $Temp_{max}$



(c) Lethbridge, AB on $Temp_{max}$

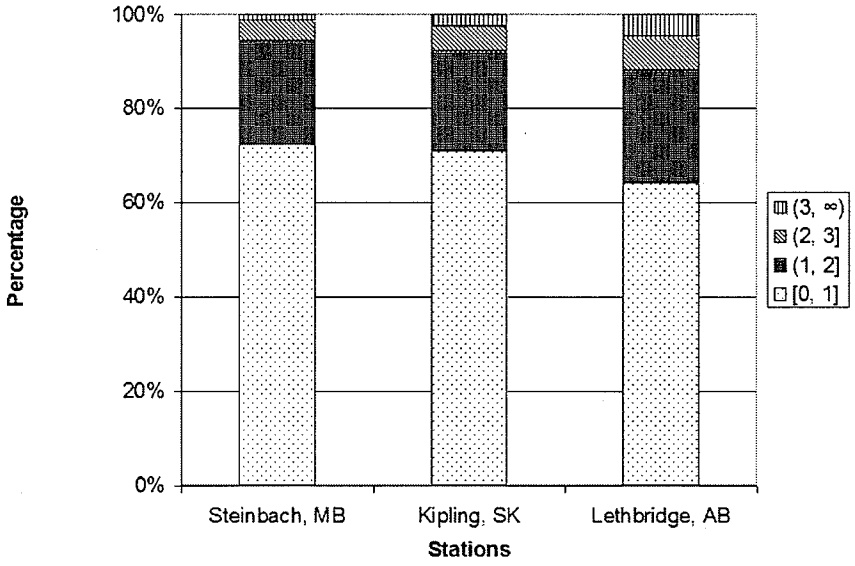
Figure 4.4: Histogram of Absolute Errors for $Temp_{max}$ (Evaluation 4.7).

For all the histogram plots, we expect to obtain a right-skewed histogram, which would indicate that the deviation of the estimates from the observed value is small. Specifically, we would like to obtain the highest frequencies (i.e., tallest histogram bars) on small x-axis values, which would indicate that the majority of the predicted values had a small difference from the observed values and thus, a good estimation performance.

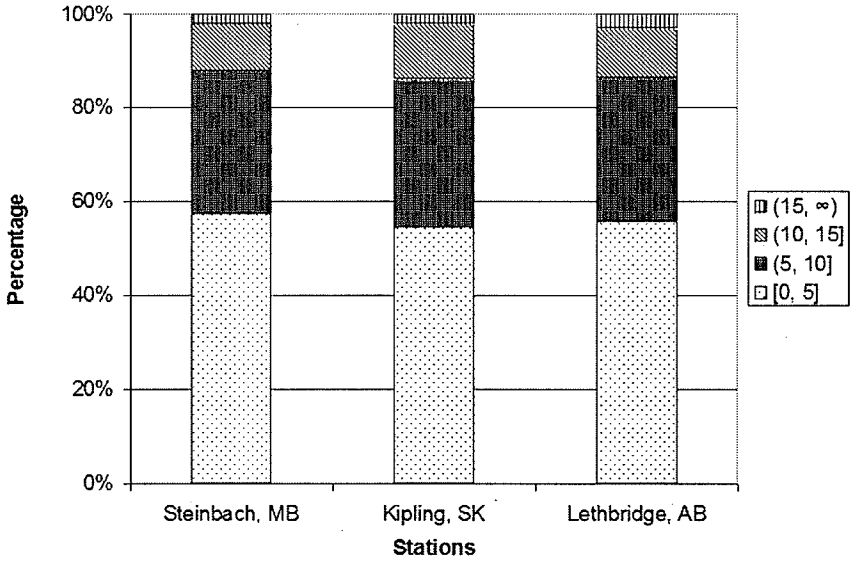
For all the cumulative bar plots, we wish to obtain the greatest proportion of observed absolute errors in the lowest error magnitude interval.

Evaluation 4.7 ($Temp_{max}$ histogram) For $Temp_{max}$ (Figure 4.4), both *spatial* and *temporal estimates* showed right-skewed histograms for all three representative stations. The majority of the *temporal* estimates (left side histograms) deviated in the range of $[0.0°C, 5.0°C]$ while the majority of *spatial* estimates (right side histograms) deviated between $[0.0°C, 1.0°C]$ for the Prairie stations (`Steinback` and `Kipling`) and $[0.0°C, 2.0°C]$ for the mountainous station (`Lethbridge`). *Spatial* estimates for the prairie stations were more uniform registering lower standard deviation for their absolute errors compared to the estimates of the station in the mountainous region. Conversely, there were no remarkable differences observed in the distribution of the *temporal* estimates for prairie vis-à-vis mountainous stations.

Evaluation 4.8 ($Temp_{max}$ cumulative bar plot) Figure 4.5(a) shows the cumulative bar plot for $Temp_{max}$ spatial estimates. From the bar plots, it can be seen that the spatial estimates did very well with 80% of the estimates for the entire 30-year period lying within 2°C of the actual observed values. Furthermore, we can conclude that the spatial estimation method calculated slightly closer estimates in the prarie stations (e.g., `Steinbach` and `Kipling`), with a slightly higher percentage of errors falling just within 1°C of the actual observed value compared to the mountainous station (`Lethbridge`).



(a) Spatial Estimates



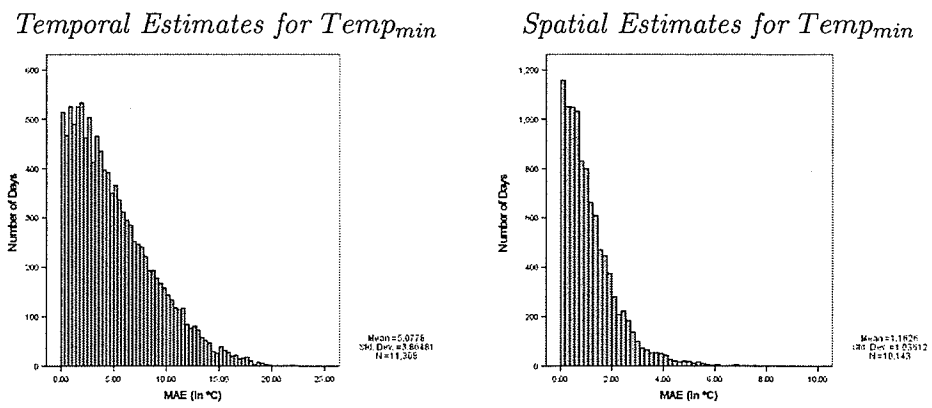
(b) Temporal Estimates

Figure 4.5: Cumulative Bar Plot of the Absolute Deviation between the Observed and Estimated Values of $Temp_{max}$ (Evaluation 4.8).

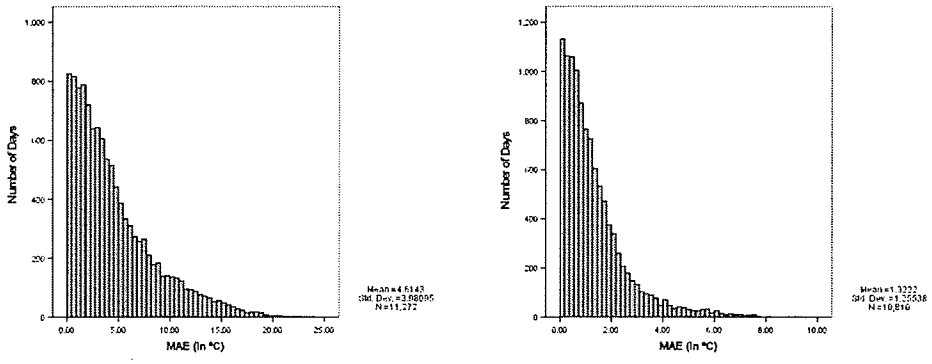
Temporal estimates for $Temp_{max}$ did fairly well with the majority of the errors lying within 5°C of the actual values observed (Figure 4.5(b)). The bar plots also reveal that no remarkable difference could be seen in the accuracy of the estimates with respect to the location of the station. This is so as all the stations returned approximately the same percentage of errors regardless of where they were located (i.e., whether they are located in the prairie plains or in mountainous regions).

Finally, from our cumulative bar plots, we conclude that our temporal estimation method can predict within 5°C of the actual values observed approximately 58% of the time.

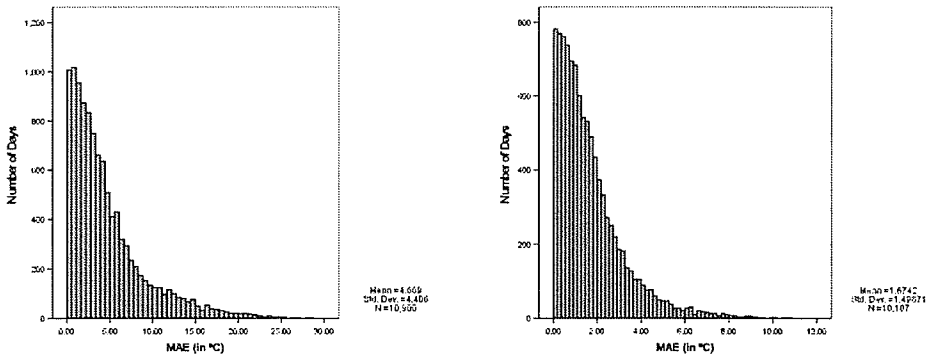
Evaluation 4.9 ($Temp_{min}$ histogram) For $Temp_{min}$ (Figure 4.6), the same right-skewed histograms resulted for both the *spatial* and *temporal* estimates for all three representative stations. Compared to $Temp_{max}$, $Temp_{min}$ estimates for both methods received higher absolute error values. The majority of the *temporal* estimates varied in the range of $[0.0°C, 10.0°C]$ whereas the majority of *spatial* estimates deviated by less, within only 2°C of the actual values for the prairie stations (`Steinback` and `Kipling`) and 4°C for the mountainous station (`Lethbridge`). Unlike $Temp_{max}$, there were no remarkable differences in the distribution of both the *temporal* and *spatial* estimates between the prairie and mountainous stations.



(a) Steinbach, MB on $Temp_{min}$



(b) Kipling, SK on $Temp_{min}$



(c) Lethbridge, AB on $Temp_{min}$

Figure 4.6: Histogram of Absolute Errors for $Temp_{min}$ (Evaluation 4.9).

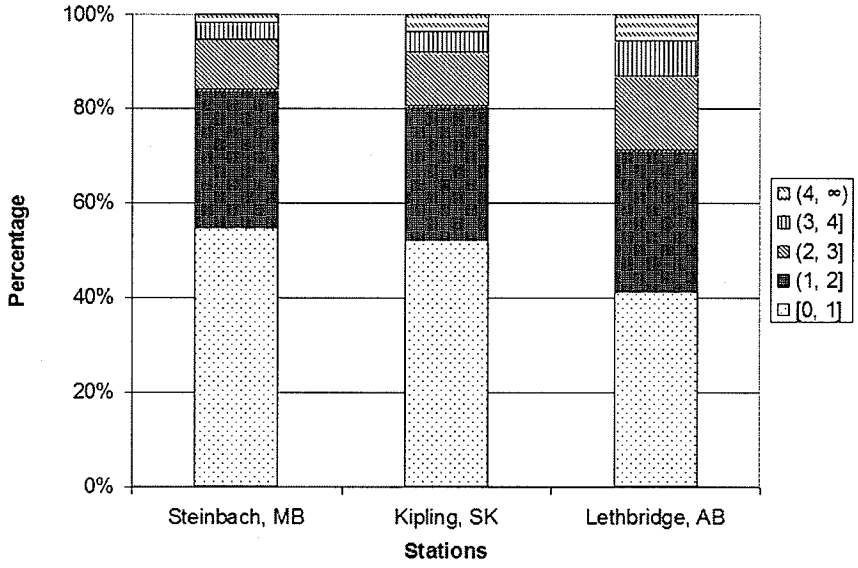
Evaluation 4.10 ($Temp_{min}$ cumulative bar plot) From Figure 4.7(a), our cumulative bar plots indicate that the spatial estimation method can estimate within 2°C of the actual observed value 75% of the time. The spatial estimation method performed better in the prairies (i.e, `Steinbach` and `Kippling`) as compared to the mountanous regions. This is because the stations in the prairies received a greater percentage of errors with lower magnitude (i.e., approximately 55% of errors for `Steinbach` and `Kipling` fell in the interval [0.0°C, 1.0°C] vs. 40% for `Lethbridge` for the same error magnitude).

Similar to $Temp_{max}$, temporal estimates for $Temp_{min}$ did fairly well with the majority of the errors lying within 5°C of the actual values observed (Figure 4.7(b)). Temporal estimates did well on mountainous regions, with the estimates at `Lethbridge` station receiving a greater percentage of smaller magnitude errors compared to its prairie station counterparts in `Steinbach` and `Kipling`. From our cumulative bar plots, we can conclude that our temporal estimation method can predict within 5°C of the actual values approximately 58% of the time.

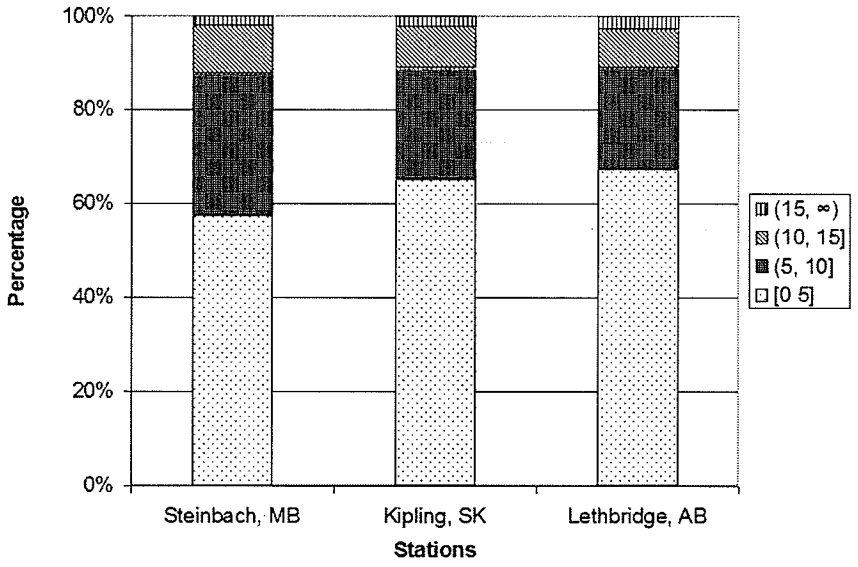
Table 4.2 presents a summary of our histogram and cumulative bar plot results.

Table 4.2: Summary Table for Histogram and Cumulative Bar Plots.

	Temporal Method	**Spatial Method**
$Temp_{max}$	prediction within 5°C of actual values 58% of the time	prediction within 2°C of actual values 80% of the time
$Temp_{min}$	prediction within 5°C of actual values 58% of the time	prediction within 2°C of actual values 75% of the time



(a) Spatial Estimates



(b) Temporal Estimates

Figure 4.7: Cumulative Bar Plot of the Absolute Deviation between the Observed and Estimated Values of $Temp_{min}$ (Evaluation 4.10).

4.4 Experiment Set 3: Accuracy of Estimates–Pearson Correlation Coefficient (R^2) & Nash-Sutcliffe Coefficient of Efficiency (E) Tests

To evaluate the "goodness-of-fit" of our estimates, we used both the *Pearson correlation coefficient* (R^2) [Moo04] and *Nash and Sutcliffe efficiency coefficient* (E) [NS70] in the pairwise comparison between the original and the predicted values for each station. The *Pearson correlation coefficient* R^2 is given by:

$$R^2 = \frac{\left[\sum_{j=1}^{N} \left(\widehat{x}_j - \bar{\widehat{x}}\right)\left(x_j - \bar{x}\right)\right]^2}{\sum_{j=1}^{N} \left(\widehat{x}_j - \bar{\widehat{x}}\right)^2 \sum_{j=1}^{N} \left(x_j - \bar{x}\right)^2} \tag{4.7}$$

The *coefficient of efficiency* E is given by:

$$E = 1 - \frac{\sum_{j=1}^{N} \left(\widehat{x}_j - x_j\right)^2}{\sum_{j=1}^{N} \left(\widehat{x}_j - \bar{x}\right)^2} \tag{4.8}$$

For both of these equations, x_j is the original value at date j, $\widehat{x}_j$ is the corresponding estimate at date j, $\bar{x}$ is the arithmetic mean of the original values x_j, and $\bar{\widehat{x}}$ is the arithmetic mean of the estimates $\widehat{x}_j$ from dates 1 to N. Ideally, we would like to have values of R^2 and E close to 1.

A value of E equal to 1 indicates that all the estimated values exactly correspond to their original values. On the other hand, low values of E indicate large deviations from the original observations. Lastly, a negative E, indicates that it is better to use the mean

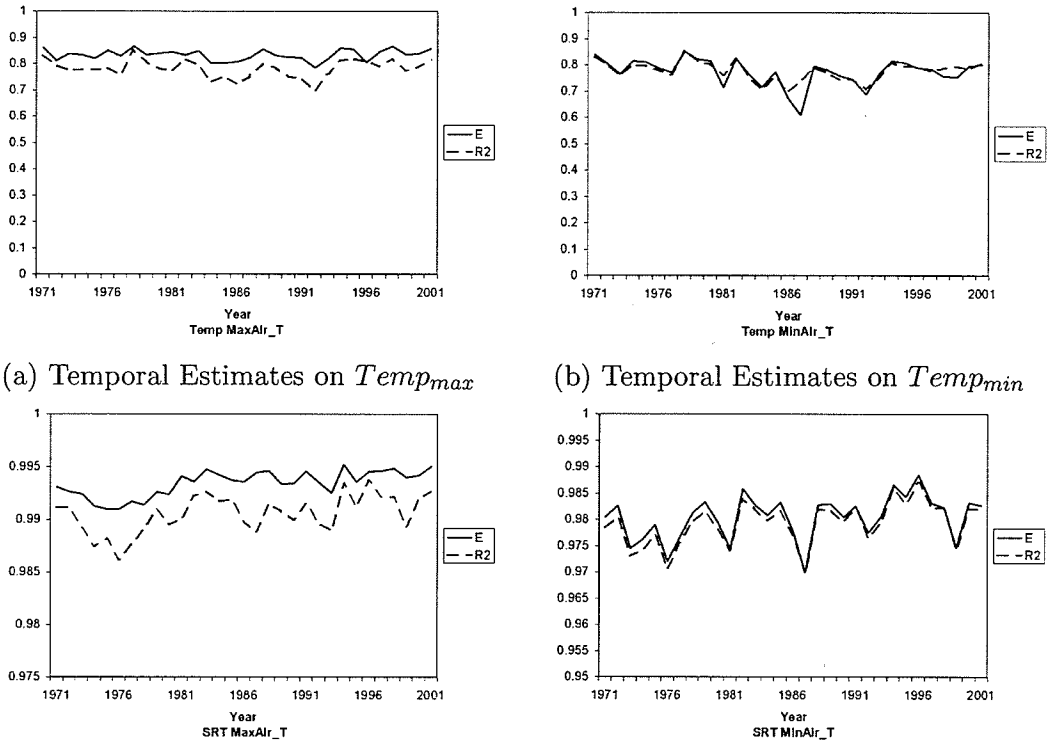
Table 4.3: Average correlation coefficient (R^2) and coefficient of efficiency (E) between the original and predicted values for each station for a 30-year dataset (Evaluation 4.11).

	Temporal		**Spatial**	
	$Temp_{max}$	$Temp_{min}$	$Temp_{max}$	$Temp_{min}$
Avg R^2	0.780	0.777	0.990	0.979
Max R^2	0.910	0.908	0.999	0.997
Min R^2	0.449	0.168	0.854	0.839
Avg E	0.833	0.774	0.993	0.980
Max E	0.910	0.905	0.999	0.997
Min E	0.449	0.288	0.911	0.842

$\bar{x}$ as an estimator. In our results, we consider E values 0.75 and above as having a strong agreement between the estimates and the actual observation.

Evaluation 4.11 (E and R^2 tabular results) Table 4.3 presents a summary of E and R^2 values. Both the temporal and spatial estimation methods registered high E values (i.e., E values close to 1). This indicates high values for our E and implies that our estimates agreed well with the actual observed values. For example, the *temporal* estimates for $Temp_{max}$ did well, registering an average E coefficient value of 0.833. Moreover, the *spatial* estimate almost did a perfect estimate of the actual value, with the maximum and minimum E coefficients for $Temp_{max}$ registering at 0.999 and 0.911 respectively.

Evaluation 4.12 (E and R^2 time plots) To identify any trend in the coefficients over the years, we plotted the average R^2 and E values for all stations on a yearly basis for both the maximum and minimum temperature attributes (see Figure 4.8). The *temporal* estimates for $Temp_{max}$ registered very high E values, with E fluctuating between 0.8 and 0.9 for the entire 30-year period (see Figure 4.8(a)). The same trend was observed for the Pearson correlation coefficient R^2, with its values mostly following those of E. Temporal estimate results for $Temp_{min}$ registered satisfactorily high E and R^2 values with the E values registering around the range [0.65°C, 0.90°C]. For $Temp_{min}$, however, the changes in



(a) Temporal Estimates on $Temp_{max}$

(b) Temporal Estimates on $Temp_{min}$

(c) Spatial Estimates on $Temp_{max}$

(d) Spatial Estimates on $Temp_{min}$

Figure 4.8: Coefficient of Efficiency E and R^2 plots for temporal and spatial estimates (Evaluation 4.12).

the values of E were more drastic (see Figure 4.8(b)), where we observed the least accurate estimates during the 1986-1988 period.

Spatial estimates for $Temp_{max}$ registered extremely high E values, with E registering around the interval $[0.990°C, 0.995°C]$ (see Figure 4.8(c)). The lowest value was observed during the late 1970s, which then increased over the next few years. The Pearson correlation coefficient R^2 has a slightly lower value but followed the trend of E. The E values for $Temp_{min}$ (see Figure 4.8(d)) displayed greater heterogeneity, with downward spikes occurring during the 1986-1988 period, similar to what we observed for the temporal estimates. Despite this, the spatial estimates still performed extremely well, registering values in the range $[0.97°C, 0.99°C]$.

4.5 Summary

The results of random seeding experiments show that our QC tool is able to flag erroneous data accurately. The high *F-measure* values empirically confirmed the tool's effectiveness in flagging erroneous observations.

Our evaluation of the accuracy of estimates gave very good results. The experiments show that our QC tool is able to calculate estimates as near as 2°C of the actual values 80% of the time. Our tests also showed that between *spatial* and *temporal* estimation methods, the *spatial* method calculated estimates that, in majority of cases, deviate less from the observed values than those using the temporal method.

Tests on E and R^2 values reveal a strong measure of "goodness-of-fit" between our estimates and the actual observed values. As with the previous test, we noticed that the spatial estimation method performed better than the temporal estimation method. However, this result does not imply that the estimates from the temporal method are inaccurate and entirely useless. As indicated by the values of the efficiency coefficient E, our temporal method still obtained a high value.

Since the spatial method returned better MAE, RMSE, E, and R^2 values in all evaluation cases, we select the *spatial estimation method* as our first choice as an estimator. Despite the fact that the *spatial* method gave more accurate estimates than the *temporal* method, there are cases where a spatial estimate is just difficult or impossible to obtain. These cases usually occur in isolated weather stations with no immediate neighbours (e.g., stations in Thompson and Churchill in northern Manitoba), and thus insufficient qualified stations would be available to compute a spatial estimate. In situations such as these, we may have historical data available for the station, and therefore, temporal estimates can be calculated and would still be useful. Hence, we have *temporal estimates* as an alternative if a spatial estimate is unavailable or difficult to calculate.

Chapter 5

Challenges & Solutions in Designing and Developing the Multi-Layer QC Tool

So far, we have described our multi-layer tool for detecting abnormal weather observations. We discussed each implementation layer and how the quality control algorithms in each layer works. We have also discussed the procedures taken by our tool in calculating an estimate to fill-in flagged or missing data. Finally, in Chapter 4, we assessed our tool's capability in flagging erroneous observations and calculating accurate estimates.

In this chapter, we focus on the implementation details and discuss the challenges we encountered in developing our multi-layer quality control tool. We also present the accompanying solutions that we applied to each problem.

5.1 Manitoba Agriculture and Its Operational Environment

In the history of agriculture, weather has proved to be the most important factor affecting crop production. Agricultural yields and production are highly correlated to the prevailing weather conditions in a specific area. Management decisions in farming, such as (i) when to best administer pesticides and fungicides, (ii) which particular crop variety to cultivate, and (iii) how to plan for irrigation during the growing season, are all influenced by the weather. As such, it is crucial for farmers to be guided accordingly on their management decisions for a favourable yield on their crops to make the most out of the growing season.

Rather than entirely relying on personal insights and past farming experiences, both crop scientists and agro-meteorologists have developed agricultural models that assist farmers in making management decisions regarding their crops. Two such decision support models are the potato Disease Severity Value (DSV) model [Wal62] and the corn Crop Heat (CH) model [GR58]. The first model, the potato DSV, helps farmers determine when is the best time to administer fungicides to prevent the potato late blight disease, a serious yield-debilitating fungal disease of potatoes that has, in the past, caused widespread famines in Europe. Similarly, the corn CH model guides farmers on the best corn variety to be cultivated during the growing season in their area. Both these models are dependent on weather data such as temperature, precipitation and relative humidity for their calculation.

Being a traditionally agricultural province, Manitoba Agriculture, Food and Rural Initiatives (MAFRI) operates a real-time network of stations which gathers weather data throughout the province of Manitoba. From the data gathered from its weather network, MAFRI operates both the potato DSV model and the corn CH model, and provides general monitoring information and decision support models to agricultural producers in Manitoba. This network is Internet-based and phone-based, which facilitates the dissemination of site-

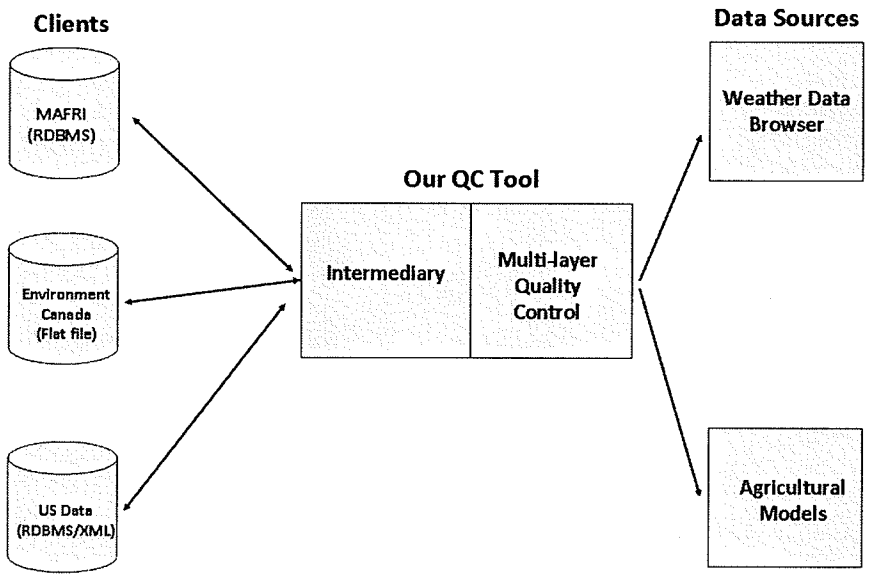


Figure 5.1: Architecture of our QC tool with the data intermediary module and its relation to the diversified, inter-format data sources shown.

specific weather-based recommendations, and allows it to maximize its reach even in the far-flung regions of Manitoba. This is specifically valuable in the case of Manitoba, with its large geographic expanse resulting in remote agricultural regions. Since farmers base their expensive management decisions on the results from these models, it is crucial for MAFRI to provide reliable and accurate model forecasts and guidance to the farmers—their effective stakeholders. A key requirement to have an accurate model is to have reliable and error-free data. Uncertainty in the data brings about doubts on the derived outputs of the model and thus may compromise the decisions made by farmers. In such scenarios, quality control is necessary to scrutinize the input data and eventually ensure the accuracy of the derived advice given to farmers from these decision support models.

5.2 Implementation Challenge: Data Format Incompatibility

As with any project, we encountered various challenges during the course of implementing our quality control tool. One of the most important difficulties that we encountered was related to data integration and data compatibility issues. Recall that different stages of our implementation layers, most particularly the *temporal layer*, require a complete archive of data spanning the past several decades to perform temporal data QC. Since the Manitoba Ag-Weather Program was relatively recently established, MAFRI's current weather database did not contain all the needed datasets. For instance, it did not provide historical weather data for 30 years or longer, which are needed in temporal QC. Moreover, it did not provide enough data for ample geographical coverage to obtain an accurate spatial estimate for doing spatial QC. In short, the MAFRI weather dataset was not sufficient alone for our QC requirements. Hence, we needed to search for alternative data sources, such as Environment Canada (EC). Unfortunately, as most weather data gathering organizations are autonomous, there is very little collaboration seen between different organizations and little consistency between their data sources. Using EC data was technically difficult due to format incompatibility. Environment Canada's Canadian National Climate Data and Information Archive, which contains Canadian weather observations running from the late 1800s, can be used to fill in the required historical data. However, such an archive was implemented using a legacy flat file format (see Figure 5.2 for a snapshot from such an archive). Older databases pre-dating the RDBMS era were implemented using special purpose flat file formats. These older data storage systems were implemented when inter-platform data exchanges were not common, and thus collaborative data use between two or more different data sources was not considered in the data system design.

Figure 5.2: User interface to the flat file database for the Canadian National Climate Data and Information Archive (Source: *Canadian Daily Climate Data CD, Environment Canada*).

As a solution to this problem, we implemented an intermediary tool such as that described by Laurenson et al. [LKN02]. This intermediary tool allowed us to retrieve and use data from various sources by presenting a standardized view of the data from both the present-day RDBMS and the legacy flat file system. During data retrieval, our intermediary creates a standardized view, where individual data representation is hidden from the requesting application. For each data retrieval operation, the intermediary tool decides where to pass the query to (from among the different data sources), and converts it to native queries so that it can be executed by the receiving database. The results are then formatted and passed to the requesting application, which can read the requested data in a format that is familiar and intelligible.

5.3 Implementation Challenge: National Units Standards Differences

Another challenge encountered is the unavoidable differences in national standards brought about by the diversity in time period and geographical provenance of our archival datasets. Recall that, in the *spatial layer*, we need to have observations from surrounding stations to calculate a confidence interval in which a weighted estimate is obtained (Section 3.4). Some stations in Manitoba are almost at the US border, which necessitated the use of North Dakotan data. This creates a unit standard conflict because North Dakotan stations still record observations in Imperial units (e.g., Fahrenheit for temperature, inches for precipitation) while Manitoban stations record in metric even if the stations are just a few kilometres apart across the border. Analogously, Canada was under the Imperial standard before the 1970s, and raw data from these periods cannot just be combined and used with the present data without the appropriate pre-processing and conversion. At first consideration, this problem might seem to be trivial with a straightforward solution (i.e., unit conversion), but with the large amount of data that we are working with, a simple mix-up with different units will surely create problems which in the end will affect the integrity of the results!

As a solution, we implemented a module that automatically determines whether a standard or unit conflict occurs between datasets. A unit conversion is performed using knowledge of the time period when the data were recorded and the existing national measurement standard in effect for the station.

5.4 Summary

In this chapter, we described the implementation challenges we encountered in implementing our QC tool. For each challenge, we proposed and implemented accompanying solutions to solve them. We built a QC tool specifically for Manitoba Agriculture in its mission to improve the yield of Manitoba farmers by providing guidance on their management practice decisions.

Since we are using data from various sources and time eras, we encountered two major problems pertaining to (a) data format incompatibility and (b) national standards differences. We solved Problem (a) by implementing an intermediary that converts foreign data formats into an intelligible format that is compatible with our QC tool. We solved Problem (b) by implementing a module that automatically determines whether a standard or a unit conflict occurs between datasets and does the necessary standards conversion.

Chapter 6

Generalizing Our Multi-Layer Tool to Other Applications

In Chapter 3, we learned how our multi-layer QC tool can be applied for the quality assurance of weather data. A common theme throughout Chapter 3 was on how various logical, statistical tests and data mining techniques can be integrated and applied together for the quality control of weather data. In this chapter, we extend the purpose of our quality tool further, showing that is it not only applicable for weather but also in other diversified fields besides agro-meteorology.

In this chapter, we direct our attention to case studies on other applications where our tool could be applied. We show that the tool is generalizable to use in other application domains. We discuss other possible applications from four different fields namely, *utility consumption monitoring*, *city pest control*, *traffic engineering*, and *e-health*.

6.1 Utility Consumption Monitoring

First, our model can be applied to the monitoring of utilities consumption such as electric power, natural gas, water supply, and telecommunication services. In today's era of modern technology and service delivery, utilities form a significant part of daily consumer consumption. Utilities are considered to be universally needed services that are provided to homes and businesses. In some parts of the world, people rely on the reliability of their utility services for daily survival. This is especially true for Canada where, during the winter months, Canadians rely on the reliability of hydro power and/or natural gas services to provide heating and keep them warm despite the intense sub-zero temperatures. With this large reliance upon energy sources, it is important that only a suitable amount of energy is consumed so as to prevent wastage and to ensure efficient use. To this end, we can apply the *internal layer* of our proposed model to ensure that readings from power, gas, or water meters are valid and fall within reasonable ranges. We can apply the *temporal layer* to check the consistency of utility consumption of each household. In utilities, it is expected that consumption patterns will follow a yearly cycle. For example, we expect higher natural gas consumption during winter for heating. Moreover, we can also apply the *spatial layer* to check the consistency of utility consumption (e.g., applying the spatial regression test to monitor readings on consumer consumption of electricity or natural gas) of households in certain "groups". Consumers in a certain locality could be "grouped" together based on their past consumption and compared with each other. This grouping prevents unfair comparison (e.g., business consumers with residential ones, or consumers occupying large houses with ones owning a starter home). If there is an increase in utility consumption by one consumer, consumption by other consumers in the same "group" will be increased as well (e.g., an increase in natural gas consumption for heating due to record-

breaking low temperatures in an area). Any abnormal or outlying data can be detected by the spatial layer of our model. These data could be indications of a malfunction on the meter reading devices or a true heightened utility consumption by the consumer. For the former, the utility company could take appropriate actions such as inspecting the meter of the consumer. For the latter, the utility company could provide the consumer with energy saving tips and recommend house inspection for appropriate preventative actions such as repairs or insulation.

6.2 City-wide Mosquito Control in Winnipeg

The city of Winnipeg experiences an excessive growth of mosquito population during the spring and summer months. As a response to past human cases involving West Nile virus, Winnipeg city authorities have set-up comprehensive mosquito trap counters around the city to aid in decision-making on when it is necessary to administer chemical-based pesticide fogging. Various non-pesticide control tools are in place while mosquito levels are deemed to be in a controllable level; however to ensure public health safety, the city sometimes resorts to malathion pesticide fogging which is a constant source of friction and controversy to its residents due to malathion's potential long-term health effects.

The spatial regression test can be used in a city's mosquito surveillance and control measures for public health purposes. Spatial regression checks can be used by city administrators to formulate an integrated pest management strategy to suppress mosquito population overgrowth in areas where this is required and to ultimately reduce public health risks brought about by these disease-carrying insects.

Modelling in this application might be done as follows: mosquito traps are scattered throughout the city to monitor adult mosquito counts in a certain locality of the

Figure 6.1: City of Winnipeg mosquito trap locations (Source: *The City of Winnipeg website,* `http://www.winnipeg.ca`).

city. Under normal conditions, the mosquito count of a mosquito trap should not be very dramatically different from the count of its neighbouring traps. A statistically significant difference on counts between a specific locality and its surrounding locality sets an alarm for the city's Insect Control Branch to investigate the flagged locality for possible pesticide application or fogging.

For city authorities, knowing which district of the city is at risk for high pest count also allows efficient resource allocation considering Winnipeg's vast geographic expanse. Information on which locality has the highest risk for high trap counts will help the city in targeting high risk areas to plan public awareness campaigns and West Nile virus prevention programs.

6.3 Highway Traffic Monitoring and Law Enforcement

To ensure public road safety, various traffic law enforcement agencies have decided to install cameras for identifying vehicles violating traffic regulations, usually speed limit restrictions. For this, cameras are usually installed in various highway sections to monitor traffic density and speed limit violations.

Spatial regression testing can be applied in two instances for highway traffic monitoring. The first involves detecting highway sections where high rates of vehicle speeding occurs. We can compare camera stations on highway sections that have the same speed limit and vehicular density as one another in terms of the number of speeding vehicles caught.

Once we identify these similar stations, we can apply spatial regression testing to determine if a particular station differs significantly in the number of speeding vehicles caught with respect to other traffic cameras. Two scenarios exist on this finding, either we have malfunctioning equipment or there has been a genuine increase in violators in that particular section of the highway. We can then apply follow-up procedures—first, we can have a look into the machine, and secondly, we can station more traffic enforcers in the area for greater police visibility to ensure greater public road safety.

The second application lies in detecting outliers based on traffic density. Most of the time, we experience bottlenecks in highway interchanges because of congestion during rush hours. For this case, we can install density detectors for each of these interchanges to have 24-hour monitoring on vehicle density. We can then compare various station readings and stations having statistically different readings can be pinpointed. In return, traffic enforcers can use this finding to make adjustments on speed limits, re-route traffic, or to deploy personnel to de-congest traffic in an area and ensure smoother flow.

6.4 Elderly Behaviour Monitoring

In today's society, innovations in health care and changing demographic trends have shifted most industrialized countries population towards an older trend. Advances in medical science provide longer life expectancy and cultural and lifestyle changes have brought about lower birth rates among fertile age groups, resulting in an increased percentage of seniors in the population.

One of the fastest aging societies in the world, Japan in the 1950s had 9.3 persons of age 20–64 for every person over 65. By 2020, this ratio is forecast to be at 0.59 people under 20 for every person over 65 [Uni06]. This problem is not only confined to Japan but is also occurring in European countries as well. Spain, for example, has had elementary schools closing in rural areas due to lack of enrolment. In Italy, the high percentage of seniors serves as a challenge for social security planners and has put serious strains in the country's pension schemes.

With this demographic trend comes an increasing demand for elderly caregiver and monitoring services. However, certain elderly persons are not receptive to being dependant on caregivers and would prefer to live independently. Unfortunately, this may leave them vulnerable to accidents and eventual neglect. The procedures in our QC tool can also be used for automated elderly monitoring, allowing independent living while ensuring continued monitoring for the elderly subjects.

Temporal outlier checks can be used to develop a system to monitor the health-related behaviour of elderly people living alone. Time series data on in-house movements of subjects can be collected on a specific interval basis. Through other data mining techniques, patterns of movements can be discovered. An elderly individual's health condition can be estimated by comparing data on the duration of stays in specific rooms such as the lava-

tory and daily behavioural patterns such as getting water from a thermos with previously recorded data. If an anomalous pattern or observation is detected, family members could be informed automatically via text or telephone who can then make the necessary decision to seek the appropriate help or medical assistance.

6.5 Summary

After assessing the capabilities of our QC tool on weather data in the previous chapter, we discussed the applicability of our tool to other daily applications in this chapter. We considered four different application domains: utility consumption monitoring, city pest control, traffic engineering and e-health.

In utility consumption monitoring, the *internal layer* can ensure that readings from utility meters are within their valid ranges. With the internal layer, we can detect malfunctioning meters or abnormal consumption patterns for a specific customer. In a city's pest control program, the *spatial layer* can be used to check which area of the city has abnormally high mosquito counts and can be used to alert authorities so that appropriate insect control measures can be undertaken. We can also apply the *temporal layer* in traffic engineering to monitor cyclical patterns on the daily vehicle use of highway sections and use the patterns discovered to identify times where traffic bottlenecks are likely to occur. In the same manner, we can also use the *temporal layer* to observe behavioural changes of elderly people living alone via the detection of subject movements through sensors. As their behaviour usually follows a cycle, any significant change in this behaviour might indicate an emergency situation (i.e., elderly non-mobility) so that we can take appropriate emergency action.

In closing, we note that the variety of applications domains possible for our tool

effectively indicates its flexibility and successful generalizability not only to weather data but also in other applications as well.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Weather plays an important role in agriculture. To aid farmers, crop scientists and agro-meteorologists have devised many agricultural decision support tools to improve farm management decision practices. However, these agricultural decision support tools are weather-driven and the accuracy of these models and the success of the decisions made from them depends heavily on the quality of their input data. Thus, a clean and complete set of data is essential for their successful operation and subsequent data mining operations. Otherwise, the results may be misleading or worse, entirely useless.

Outlier detection, a data mining technique that identifies observations that are different from most other observations, is a technique that can be used for scrutinizing the quality of our weather data. While the scientific literature contains abundant outlier detection techniques for general data, the availability of work concerning the application of these techniques to the quality control of weather data remains scarce. From these issues, we identified the following research questions which we addressed in this thesis. These questions (with their respective answers) are as follows:

(a) *How can we develop an automated system for detecting abnormal weather observations using outlier detection methods to control the quality of MAFRI's weather data and enhance its agricultural decision support models?*

In this thesis, we developed a multi-layer quality control (QC) tool that integrates single-station and multi-station techniques to detect and remove abnormal observations from weather data (Chapter 3). Our QC tool enhances MAFRI's weather-driven agricultural decision support models by ensuring that each datum is accurate, timely, and consistent for the agricultural models to use. Ensuring the quality of the input data equally ensures us of an accurate final output for our intended applications, and that is our agricultural decision models.

(b) *How can we improve existing data quality control methods to enhance weather-driven decision support models?*

Our QC tool uses quality control techniques that use multiple stations in calculating an evaluation parameter (Section 3.4). In the multi-station techniques, we used a combination of statistical and data mining outlier detection techniques. Our tool uses a statistical approach based on correlation in selecting which surrounding stations to include as neighbours. Unlike past techniques, our tool does not assume that the best station to compare against is the closest station, but instead, it looks into the behavioural similarities (when it comes to measurements) among stations as a determinant in choosing which station to include in calculating the evaluation parameter. This allows our technique to perform better on the quality control of stations that are close to one another than those located in two different climatic regions due to topographical features.

(c) *How can we apply outlier detection techniques on the quality control of weather data and fill-in discarded and missing observations to maintain a complete set of data?*

Besides ensuring the quality of weather data, we provided a mechanism to replace flagged erroneous dataset (Section 3.5). With this data filling mechanism, we are able to maintain a complete time series dataset that is essential for the seamless operation of weather-driven agricultural decision support models. In calculating estimates, our tool uses two estimation methods namely *temporal* and *spatial*. Calculated parameters from the *spatial* and *temporal* layers, which are used to decide whether to flag a datum, act as a good estimate when used as a stand-alone parameter. Moreover, it can also be used to fill in flagged or missing data.

(d) *How can we integrate incompatible weather data formats into compatible ones for collaborative use?*

In our discussions, we documented the problems that we encountered in implementing our QC tool (Chapter 5). As we used data from diversified geographic provenance, we had data that were encoded in various formats and used different national standards which made them incompatible with one another. As a solution, we devised a data intermediary that allows data from various sources and using different national standards to be retrieved and used.

(e) *How can we test the effectiveness of our proposed methods?*

Our experiments showed that our quality control tool is able to flag erroneous data accurately and is able to calculate accurate estimates to fill in flagged and missing data (Chapter 4). In all the stations tested in our experiments, our tool registered a 70-80% range of successfully flagging seeded errors. Results from the accuracy of estimates test show that our QC tool calculates very accurate estimates with the

ability to predict as near as 2°C of the actual values 75–80% of the time. While it can be said from our results that the *spatial* estimation method outperformed the *temporal* estimates, it does not mean that the *temporal* method is unsuccessful. As the *spatial* method calculates more accurate estimates, we designate this method as our first choice and have the *temporal* method as an alternative when a *spatial* estimate is difficult or impossible to obtain.

(f) *How can we extend our tool from weather data to generalize its applicability to other diversified application domains?*

Finally, we presented a number of real-life applications in which our tool can be generalized and applied. We discussed how our tool can be generalized to applications such as utility consumption monitoring, city pest control program, traffic engineering, and telemedicine. Each layer in our QC tool can be extended for successful use in each of the mentioned application domains. The myriad of possible application domains for our tool effectively indicates its flexibility and successful generalizability not only to weather data but also to other applications as well.

To summarize, our key contributions in this thesis are:

1. We developed a multi-layer quality control tool that enhances weather-driven agricultural decision support models. We used both *spatial* and *temporal* checks in the quality control algorithms that compares a station's measurement with multiple surrounding stations and compares a station's present measurement with its previous archival measurement, respectively.

2. We developed techniques to maintain a complete time series weather dataset through the calculation of estimates to fill in flagged or missing data. This is specifically helpful

for agro-meteorological decision models that can only work when there is a complete time series dataset input.

3. We assessed the techniques used in our tool to show that the tool is not only applicable for weather data but in other diversified applications as well.

7.2 Future Work

For future work, we plan to investigate on the following: (i) data compatibility and integration issues brought about by new generation extensible mark-up language (XML) weather datasets and (ii) the empirical evaluation of the effect of our QC tool on agricultural decision models.

First, we plan to investigate the problems associated with the eventual move of weather data to XML. XML in recent years has gained popularity as a method for representing data. Because of its popularity and the perceived advantages it brings, national weather bureaux such as the US National Weather Service (NWS) have started providing weather data in an open-access XML format known as the National Digital Forecast Database (NDFD). A variety of weather parameters are contained in the NDFD XML weather database such as maximum, minimum, hourly and dew point temperatures, precipitation amount, wind direction and speed. With this structure, anyone with access to the NWS can retrieve weather information in XML format. NWS currently has its own XML language known as the Digital Weather Mark-up Language (DWML) and issues its own schema.

XML is an Internet technology created to facilitate data exchange and storage. It is a general purpose mark-up language which supports the sharing of structured data in various mediums specifically the Internet. The standard is an open one and is recommended

by the World Wide Web Consortium (W3C). For data retrieval, querying and processing, it is supported by a suite of technologies such as XSLT, XPath and XQL.

There are various benefits in having weather databases represented in XML. First, representing weather data in XML makes the data available not only in machine-readable but also in a human readable self-describing form. Unlike existing databases, this makes the data representation independent of any computing platform and database vendor. Moreover, XML has existing technologies that accompany it which makes integration of the data in HTML and web applications easier. This is important for weather databases, where the majority of the dissemination and distribution to the public is done via the Internet. Having XML will open instant weather data access to portable electronics equipment such as mobile phones which is of special interest to farmers working in the field. Together with the advancing capabilities of mobile devices, it opens the way to have important agricultural management decision models readily accessible on the spot by the farmers while working in the field.

As weather databases gear toward XML as the standard, another set of compatibility issues will arise with present and past data representations during collaborative data exchange and usage. However, we think that unlike the case of legacy flat file databases, portability issues will be kept to a minimum as conversion tools are readily available.

Finally, in this thesis, we determined through our experiments that our tool is capable of successfully identifying erroneous observations. As future work, we plan to extend this evaluation from our QC tool to the agricultural decision models themselves, as they are the main "consumers" of the quality controlled data. As an evaluation, we can compare the quality of the outputs of the agricultural decision models when using quality controlled data versus non-quality controlled data. To see the concrete benefits of using our QC tool to farmers, we can evaluate the effects of using the decisions based on the

agricultural models to farmers's yield. We can set up an experimental farm that bases its management decisions on agricultural models using non-quality controlled data as a "control" and a farm that bases its management decisions from quality controlled data as the "treatment" and compare the crops' general yield and health in each farm to determine if there is a significant change between the two.

Bibliography

[BBB06] M. Benedikt, P. Bohannon, and G. Bruns. Data cleaning for decision support. In *Proceedings of the International Very Large Databases (VLDB) Workshop on Clean Databases*, pages 1727–1732, Seoul, Korea, September 2006.

[BP85] D. P. Ballou and H. L. Pazer. Modeling data and process quality in multi-input, multi-output information system. *Management Science*, 31(2):150–162, 1985.

[CCV07] S. Cateni, V. Colla, and M. Vannucci. A fuzzy logic-based method for outlier detections. In *Proceedings of the International Conference on Artificial Intelligence and Applications (IC-AI)*, pages 605–610, Las Vegas, NV, USA, June 2007.

[CFG+07] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: consistency and accuracy. In *Proceedings of the 33rd International Conference on Very Large Databases (VLDB)*, pages 315–326, Vienna, Austria, September 2007.

[CLK08] D. Chen, C. T. Lu, and Y. Kou. On detecting spatial outliers. *Geoinformatica*, 12(4):455–475, 2008.

[DG93] L. Davies and U. Gather. The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792, 1993.

[DHJ+04] P. C. Doraiswamy, J. L. Hatfield, T. J. Jackson, B. Akhmedov, J. Prueger, and A. Stern. Crop condition and yield simulations using Landsat and MODIS. *Remote Sensing of Environment*, 92(4):548–559, 2004.

[DJ04] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning.* Wiley-Interscience, Hoboken, NJ, USA, 2004.

[EAP+02] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. J. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*, pages 78–100, 2002.

[EPD+00] J. K. Eischeid, P. A. Pasteris, H. F. Diaz, M. S. Plantico, and N. J. Lott. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *Journal of Applied Meteorology*, 39(9):1580–1591, 2000.

[GDH+04] S. Goddard, J. Deogun, S. Harms, et al. A geospatial decision support for drought risk management. In *Proceedings of the 2004 Annual National Conference on Digital Government Research*, pages 35–37, Seattle, WA, USA, May 2004.

[GR58] E. C. Gilmore and J. S. Rogers. Heat units as a method of measuring maturity in corn. *Journal of Agronomy*, 50:611–615, 1958.

[Haw80] D. M. Hawkins. *Identification of Outliers. Monographs on Applied Probability and Statistics.* Chapman & Hall, Boca Raton, FL, USA, 1980.

[HGS+05] K. G. Hubbard, S. Goddard, W. D. Sorensen, N. Wells, and T. T. Osugi. Performance of quality assurance procedures for an applied climate information system. *Journal of Atmospheric and Oceanic Technology*, 22(1):97–106, 2005.

[HH02] M. Helfert and C. Herrmann. Proactive data quality management for data warehouse systems: a metadata based data quality system. In *Proceedings of the International Conference on the Design and Management of Data Warehouses (DMDW)*, pages 97–106, Toronto, ON, Canada, May 2002.

[JFM+02] C. Jacobsson, U. Fredriksson, M. Moe, L. Andresen, E. Hellsten, P. Rissanen, T. Palsdottir, and T. Arason. Quality control of meteorological observations: automatic methods used in the Nordic countries. In F. Vejen, editor, *Report 8/2002 KLMIA, Norwegian Meteorological Institute, Norway*, 2002.

[KC04] R. Kimball and J. Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data.* John Wiley & Sons, Hoboken, NJ, USA, 2004.

[KS06] C. Kondragunta and K. Shrestha. Automated real-time operational rain gauge quality-control tools in NWS hydrologic operations. In *Proceedings of AMS Hydrology*, 2006.

[KY05] S. Karatas and L. Yalcin. Data quality management. In *Proceedings of the WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (TECO)*, Bucharest, Romania, May 2005.

[Lan96] J. Lanzante. Resistant, robust and non-parametric technique for the analysis of climate data: theory and examples including applications to historical radiosonde station data. *International Journal of Climatology*, 16(11):1197–1226, 1996.

[LB08] C. K.-S. Leung and D. A. Brajczuk. Efficient mining of frequent itemsets

from data streams. In *Proceedings of the 25th British National Conference on Databases (BNCOD)*, pages 2–14, Cardiff, UK, July 2008.

[LCK03] C.-T. Lu, D. Chen, and Y. Kou. Algorithms for spatial outlier detection. In *Proceedings of the 19th International Conference on Data Engineering (ICDE)*, pages 597–600, Bangalore, India, March 2003.

[LEBQ07] Y. Lassoued, M. Essid, O. Boucelma, and M. Quafafou. Quality-driven mediation for geographic data. In *Proceedings of the 33rd International Conference on Very Large Databases (VLDB) Workshop on Quality in Databases*, pages 27–38, Vienna, Austria, September 2007.

[LEK+03] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining (ICDM)*, Montreal, QC, Canada, June 2003.

[LHL08] J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: a partition-and-detect framework. In *Proceedings of the 24th International Conference on Data Engineering (ICDE)*, pages 140–149, Cancun, Mexico, April 2008.

[LIC08] C. K.-S. Leung, P. P. Irani, and C. L. Carmichael. FIsViz: a frequent itemset visualizer. In *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 644–652, Osaka, Japan, May 2008.

[LK06] C. K.-S. Leung and Q. I. Khan. DSTree: A tree structure for the mining of frequent sets from data streams. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, pages 928–936, Hong Kong, China, December 2006.

[LKH05] C. K.-S. Leung, Q. I. Khan, and T. Hoque. CanTree: a tree structure for efficient incremental mining of frequent patterns. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)*, pages 274–281, Houston, TX, USA, November 2005.

[LKN02] M. R. Laurenson, M. T. Kiura, and S. Ninomiya. Providing agricultural models with mediated access to heterogeneous weather databases. *Applied Engineering in Agriculture*, 18(5):617–625, 2002.

[LKS05] A. Lazarevic, V. Kumar, and J. Srivastava. Intrusion detection: a survey. In *Managing Cyber Threats: Issues, Approaches and Challenges*, pages 19–80, 2005.

[LKZC07] C.-T. Lu, Y. Kou, J. Zhao, and L. Chen. Detecting and tracking regional outliers in meteorological data. *Information Sciences*, 177(8):1609–1632, 2007.

[LLN03] L. V. S. Lakshmanan, C. K.-S. Leung, and R. T. Ng. Efficient dynamic mining of constrained frequent sets. *ACM Transactions on Database Systems*, 28(4):337–389, 2003.

[LMB08] C. K.-S. Leung, M. A. F. Mateo, and D. A. Brajczuk. A tree-based approach for frequent pattern mining from uncertain data. In *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 653–661, Osaka, Japan, May 2008.

[LMN07a] C. K.-S. Leung, M. A. F. Mateo, and A. J. Nadler. CAMEL: an intelligent computational model for agro-meteorological data. In *Proceedings of the 6th International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1960–1965, Hong Kong, China, August 2007.

[LMN07b] C. K.-S. Leung, M. A. F. Mateo, and A. J. Nadler. An effective multi-layer model for controlling the quality of data. In *Proceedings of the 12th International Database Engineering & Applications Symposium (IDEAS)*, pages 28–36, Banff, AB, Canada, September 2007.

[LNM02] C. K.-S. Leung, R. T. Ng, and H. Mannila. OSSM: a segmentation approach to optimize frequency counting. In *Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE)*, pages 583–592, San Jose, CA, USA, February 2002.

[LTB06] C. K.-S. Leung, R. K. Thulasiram, and D. A. Bondarenko. An efficient system for detecting outliers from financial time series. In *Proceedings of the 23rd British National Conference on Databases (BNCOD)*, pages 190–198, Belfast, Northern Ireland, UK, July 2006.

[MDK93] T. McKee, N. Doesken, and J. Kleist. The relationship of drought frequency and duration to time scales. In *Proceedings of the American Meteorological Society Conference on Applied Climatology*, pages 179–184, 1993.

[ML08a] M. A. F. Mateo and C. K.-S. Leung. CHARIOT: a comprehensive data integration and quality assurance model for agro-meteorological data. In *Proceedings of the 13th International Conference on Database Systems for Advanced Applications (DASFAA) Workshop on Data Quality in Collaborative Information Systems*, New Delhi, India, March 2008.

[ML08b] M. A. F. Mateo and C. K.-S. Leung. Design and development of a prototype system for detecting abnormal weather observations. In *Proceedings of the 1st*

Canadian Conference on Computer Science & Software Engineering (C3S2E), pages 45–59, Montreal, QC, Canada, May 2008.

[Moo04] D. Moore. *The Basic Practice of Statistics.* W. H. Freeman and Company, New York, NY, 2004.

[NHZC07] K. Niu, C. Huang, S. Zhang, and J. Chen. ODDC: outlier detection using distance distribution clustering. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Workshop on Emerging Technologies in Knowledge Discovery and Data Mining*, pages 332–343, Nanjing, China, May 2007.

[NS70] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models. *Journal of Hydrology*, 10:282–290, 1970.

[Red01] T. C. Redman. *Data Quality: The Field Guide.* Digital Press, Woburn, MA, USA, 2001.

[Ree92] T. Reek. A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bulletin of the American Meteorological Society*, 73(6):753–762, 1992.

[RL03] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection. Wiley Series in Probability and Statistics.* John Wiley & Sons, Hoboken, NJ, USA, 2003.

[Sco03] D. W. Scott. Partial mixture estimation and outlier detection in data and regression. *Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology*, 2003.

[SFA+00] M. A. Shafer, C. A. Fiebrich, D. S. Arndt, S. E. Fredrickson, and T. W. Hughes. Quality assurance procedures in the Oklahoma mesonetwork. *Journal of Atmospheric and Oceanic Technology*, 4:435–453, 2000.

[SLZ02] S. Shekhar, C. T. Lu, and P. Zhang. Detecting graph-based spatial outliers . *International Journal of Intelligent Data Analysis (IDA)*, 6(5):451–468, 2002.

[Uni06] United Nations. *World Population Prospects.* Department of Economics and Social Affairs, United Nations, New York, NY, 2006.

[Wal62] J. R. Wallin. Summary of recent progress in predicting the late blight epidemics in United States and Canada. *American Potato Journal*, 39:306–312, 1962.

[WGH04] N. Wells, S. Goddard, and M. Hayes. A self-calibrating Palmer drought severity index. *Journal of Climate*, 17(12):2335–2351, 2004.

[WHA+08] D. J. Weston, D. J. Hand, N. M. Adams, C. Whitrow, and P. Juszczak. Plastic card fraud detection using Peer Group Analysis. *Advances in Data Analysis and Classification*, 2(1):45–621, 2008.

[Yan06] T. Yang. Neural networks for solving on-line outlier detection problems. In *Proceedings of the 2nd International Symposium on Neural Networks (ISNN)*, pages 451–456, Chengdu, China, May 2006.

[YSZ07] L. Yang, B. Shi, and X. Zhang. A new approach to outlier detection. In *Proceedings of the 2007 International Conference on Computational Science*, pages 615–620, Beijing, China, 2007.

[ZC06] Y. Zhuang and L. Chen. In-network outlier cleaning for data collection in sensor networks. In *Proceedings of the International Very Large Databases (VLDB) Workshop on Clean Databases (CleanDB)*, Seoul, Korea, September 2006.

[ZSGL07] K. Zhang, S. Shi, H. Gao, and J. Li. Unsupervised outlier detection in sensor networks using aggregation tree. In *Proceedings of the 3rd International Conference on Advanced Data Mining and Applications (ADMA)*, pages 315–326, Xi'an, China, August 2007.