

SPATIAL-TEMPORAL OUTLIER AND EVENT DETECTION IN WIRELESS SENSOR NETWORKS

ALI AMIDI
March, 2013

SUPERVISORS:

Dr. N.A.S. Hamm
Dr. Ir. N. Meratnia



SPATIAL-TEMPORAL OUTLIER AND EVENT DETECTION IN WIRELESS SENSOR NETWORKS

ALI AMIDI
Enschede, The Netherlands, March, 2013

Thesis submitted to the Faculty of Geo-information Science and Earth
Observation of the University of Twente in partial fulfilment of the requirements
for the degree of Master of Science in Geo-information Science and Earth
Observation.
Specialization: GFM

SUPERVISORS:

Dr. N.A.S. Hamm
Dr. Ir. N. Meratnia

THESIS ASSESSMENT BOARD:

Prof. Dr. Ir. A. Stein (chair)
Prof. Dr. Paul J.M. Havinga

Disclaimer

This document describes work undertaken as part of a programme of study at the Faculty of Geo-information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Many spatial phenomena are changing continuously in time and space. Thus, it is the emergence of accessing frequent, up-to-date, spatially dense measurements for monitoring and tracking. Compared to conventional earth observation data collection technologies, wireless sensor networks (WSNs) can provide continuous observations of physical phenomenon by means of dense deployment of sensor nodes. Most WSN applications require accurate, energy friendly and real-time data analysis in order to provide timely information for decision makers. Quality of data provided by WSNs is highly critical while, provided raw data may be drawn in from of a low quality and reliability level expectedly, because of the sensors inexpensive nature.

To assure quality of data obtained from WSNs, outlier detection refines the measured data and leads to the retrieval potentially useful information, called an event. There is a trade-off between accuracy and energy efficiency to design the effective event detection method. Temporal and spatial properties of occurred events are important, and thus should receive more attentive consideration.

To detect temporal outliers in the absence of a sufficient amount of historical data or in the temporary deployments, building the temporal correlation model is not usually possible thus a novel approach was proposed. The climatic observations and forecasts are utilized to identify outliers. This research evaluated the suitability of using climatic observations and climatic forecasts in WSN for outlier detection. To identify temporal outliers, patterns were utilized instead of absolute values. Experiments from the real dataset from Grand-St.-Bernard deployment in Switzerland and Italy revealed that depending on the degree of localization, forecast values can be directly applied in WSN outlier detection procedures and climatic observations can be used in building the initial temporal model. This method has the advantage of using contextual information in WSNs and was performed in an energy efficient manner.

Events were detected based on the correlation based neighbor-voting approach by integrating the temporal and spatial properties of WSNs. Events were identified where a majority of geographically correlated sensors represent temporal outliers, simultaneously. The retrieval of detailed information about spatial and temporal properties of events required the event time scale be degraded and events reclassified as faults and events in finer resolution.

Keywords

Wireless sensor networks, Temporal outlier detection, Spatial outlier detection, Event detection, MeteoSwiss forecast, MeteoSwiss observation, Patterns.

Acknowledgements

First and above all, I praise God, for providing me this opportunity and granting me the capability to proceed successfully.

I would like to express my warmest appreciation to my first supervisor, Dr. Hamm, for his constant support, scholarly advice and constructive criticisms that enhanced this work. I acknowledge and express my inestimable gratitude to my second supervisor, Dr. Meratnia whose sound advice, help and continuous encouragement contributed significantly to the completion of this study.

I would like to make a special reference to Dr. Yang Zhang, whose PhD dissertation and materials was a strong backup during my thesis. Special thanks to Tiblez for her help, despite never meeting her. I would like to thank PhD candidate Zahra Taghikhaki for sharing her experience in the topic of energy.

It is an honor for me to thank Dr. Abkar and Dr. Alesheikh for their help throughout my study in K.N.Toosi University of Technology. My faithful thanks to my adviser Dr. Khoshelham in ITC.

I would like to express my sincere gratitude to the staff of MeteoSwiss and Sensorscope for providing the data.

Last but not the least, a very deep and heart-felt thanks to my parents and my brother, for their love.

Contents

Acknowledgements	ii
List of Figures	vii
List of Tables	ix
1 Introduction	1
1.1 Motivation and problem statement	1
1.2 Research identification	4
1.2.1 Research objectives	4
1.2.2 Specific objectives and research questions	4
1.3 Innovation	5
1.4 Thesis organization	5
2 Theoretical Background	7
2.1 Outlier and event definitions	7
2.2 Spatial and temporal correlations in WSNs	8
2.2.1 Spatial modeling by geostatistics	8
2.2.1.1 Prediction	9
2.2.2 COSMO forecasting models	9
2.2.2.1 Localizing the forecast values	11
2.2.2.2 Search radius depending on model surface at station	11
2.2.2.3 Search area and preference of land points	12
2.2.2.4 Optimization of horizontal and vertical distance	12
2.3 Related work	12
2.3.1 Spatial outlier and event detection methods	12
2.3.1.1 Spatially based event definitions for outliers and events	13
2.3.2 Temporal outlier and event detection methods	13
2.3.2.1 Temporal based definitions for outliers and events	14
2.3.3 Spatial-temporal outlier and event detection methods	14
2.3.3.1 Spatial-temporal based definitions for outliers and events	16
3 Study area, Dataset description, Tools and Data pre-processing	19
3.1 Study area	19
3.2 Wireless sensor network dataset	19
3.3 MeteoSwiss dataset	21
3.3.1 MeteoSwiss observations	21
3.3.2 MeteoSwiss forecasts	23
3.4 Labeled dataset	23

3.5	Tools	24
3.6	Data pre-processing	25
3.6.1	Database management system	25
3.6.2	Data cleaning: Gross error identification	25
3.6.3	Data cleaning: Identification of dead band error	26
4	Methodology	29
4.1	Temporal outlier detection	29
4.1.1	Temporal outlier detection using patterns	32
4.1.1.1	Pattern formation	32
4.1.1.2	Similarity assessment	33
4.1.1.3	Temporal outlier detection	33
4.2	Spatial outlier detection	34
4.2.1	Spatial outlier detection by using kriging	34
4.3	Event detection	35
4.4	Re-labeling	35
4.5	Evaluation in terms of accuracy and energy	35
4.5.1	Accuracy of detected outliers	36
4.5.2	Accuracy of detected events	36
4.5.3	Evaluating the spatial model	36
4.5.4	Evaluation of complexity and energy efficiency	37
4.6	Event characterizing	37
5	Results	39
5.1	Temporal outlier detection using patterns	39
5.2	Spatial outlier detection using kriging	50
5.2.1	Evaluation of spatial model	52
5.3	Event detection	53
5.4	Results of re-labeling	55
5.5	Temporal outlier detection accuracy	56
5.6	Event detection accuracy	57
5.7	Event characterizing	59
5.8	Model complexity and energy efficiency	60
6	Discussion	63
6.1	Data pre-processing	63
6.2	Re-Labeling	64
6.3	Temporal outlier detection	64
6.4	Spatial outlier detection	66
6.5	Event detection	67
6.6	Accuracy of detected outliers and events	68
6.7	Model complexity and energy efficiency	68
7	Conclusion	69
7.1	To define outliers and events	69
7.2	To detect an event	69
7.3	To characterize an event	70
7.4	To evaluate the detected outliers and events	70
7.5	Recommendation	71

Bibliography

73

List of Figures

1.1	The organization of the thesis	5
2.1	A theatrical Gaussian model variogram	9
2.2	The three nested numerical weather prediction models of the COSMO system . .	10
3.1	The Grand St. Bernard wireless sensor network deployment	20
3.2	The illustration of ambient temperature on 2007-09-29	22
3.3	The illustration of ambient temperature on 2007-09-30	22
3.4	Illustration of ambient temperature on 2007-09-30 and 2007-09-29	22
3.5	Locations of applied datasets in the research, where M, M1, M2, M3 and M4 present the MeteoSwiss forecast grids and MO illustrates the MeteoSwiss observation station and two black clusters show the WSN small and big clusters in geographic coordinate system	23
3.6	An example of performing the plausible value check for ambient temperature, they are laid outside the (-5°C,25°C) threshold (a) before performing the plausible value check (b) after performing the plausible value check	26
4.1	The flow chart of applied methods for outlier and event detection	30
4.2	The patterns of wireless sensor network against the MeteoSwiss forecast for ambient temperature (°C) on 2007-09-29	31
5.1	The patterns of instantaneous temperature values (°C) of wireless sensor networks against the MeteoSwiss forecasts on 2007-09-30.	41
5.2	The patterns of hourly mean temperature values (°C) of wireless sensor networks against the MeteoSwiss observation on 2007-09-30.	44
5.3	The temporal outlier detection results for WSN hourly instantaneous values with respect to the MeteoSwiss forecast on 2007-09-30	48
5.4	The temporal outlier detection results for WSN hourly means values with respect to the MeteoSwiss observations on 2007-09-30	49
5.5	Illustration of variogram and the exponential fitted model where (a) variogram for 2007-09-30 and (b) variogram for 2007-09-29	50

List of Tables

3.1	Sensors specifications	21
3.2	Gross errors found using plausible value check for ambient temperature of nodes [25–28–29–31–32]	26
3.3	Dead band caused errors found using plausible rate of change for ambient temperature on nodes [25–28–29–31–32]	27
5.1	The hourly instantaneous values of WSN temperature observations on 2007–09–30 with respect to the specifications of MeteoSwiss forecasts (°C)	40
5.2	The hourly mean values of WSN measurements on 2007–09–30 with respect to the specification of MeteoSwiss observations (°C)	42
5.3	The hourly values of MeteoSwiss forecasts and MeteoSwiss observations on 2007–09–30 (°C)	43
5.4	The slopes calculated from hourly instantaneous values of WSN measurements for 2007–09–30	45
5.5	The slopes obtained from hourly mean values of WSN measurements, MeteoSwiss forecasts and MeteoSwiss observations on 2007–09–30	46
5.6	The temporal outliers presentation for hourly instantaneous values that are detected with respect to the MeteoSwiss forecasts on 2007–09–30 where, ✓ symbol presents normal measurements	47
5.7	The temporal outliers of hourly means values that are detected with respect to the MeteoSwiss observation on 2007–09–30, where ✓ presents normal measurements	47
5.8	The properties of the fitted exponential model	50
5.9	The instantaneous values of WSN temperature measurements (°C) and correspondent predictions on 2007–09–30, where “O” presents the instantaneous values of temperature and “P” presents the predicted values (°C)	51
5.10	The instantaneous values of WSN temperature measurements and correspondent predictions on 2007–09–29, where “O” presents the instantaneous values of temperature and “P” presents the predicted values (°C)	51
5.11	The correlation between wireless sensor nodes [25–28–29–31–32]	53
5.12	The result of detected events for MeteoSwiss forecasts on 2007–09–30 where, ✓ symbol presents normal measurements and ☹ symbol presents events	53
5.13	The result of detected events for MeteoSwiss observations on 2007–09–30 where, ✓ symbol presents normal measurements and ☹ symbol presents events	54
5.14	The result of re-labeling for temporal density based labeling technique on 2007–09–30	55
5.15	The result of re-labeling for temporal Mahalanobis distance-based labeling technique on 2007–09–30	55
5.16	The result of re-labeling for temporal running average-based labeling technique on 2007–09–30	55

5.17	The temporal outlier detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach of instantaneous period	56
5.18	The temporal outlier detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach hourly mean period	56
5.19	The temporal outlier detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 based on the minimum approach of re-labeled data and hourly instantaneous time period	57
5.20	The temporal outlier detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 based on the minimum approach of re-labeled data and hourly mean time period	57
5.21	The events detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 by using minimum approach of hourly period	57
5.22	The event detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach hourly instantaneous period	58
5.23	Events characteristics in fine resolution based on MeteoSwiss observations on 12:50~13:40, 2007–09–30 where, \heartsuit presents events and \textcircled{E} symbol presents faults	59
5.24	The properties of single-chip 2.4 GHz IEEE 802.15.4 compliant RF transceiver applied in wireless sensor networks (Instruments, 2007)	60
6.1	Hourly values of air humidity 2 meter above ground for MeteoSwiss forecasts and MeteoSwiss observations on 2007–09–30.	67

Abbreviations

BLUP	B est L inear U nbiased P redictor
DR	D etection R ate
ECMWF	E uropean C entre for M edium R ange W eather F orecast
FPA	F alse P ositive A larm
FPR	F alse P ositive R ate
MF	M eteoSwiss F orecast
MO	M eteoSwiss O bservation
MRF	M arkov R andom F ield
MPE	M ean P rediction E rror
RMSE	R oot M ean S quare E rror
QS	Q uarter S phere
UTC	U niversal T ime C oordinated
WMO	W orld M eteorological O rganization
WSN	W ireless S ensor N etwork

Chapter 1

Introduction

1.1 Motivation and problem statement

Many spatial phenomena are changing continuously in time and space. Thus, it is necessary to access frequent, up-to-date and spatially dense measurements for monitoring and tracking. Most of conventional earth observation data acquisition techniques provide observations that are discrete in time and not local enough. It is the emergence of new observation tools, wireless sensor networks (WSN) that can provide new options of sensing. Wireless sensor networks can observe physical phenomenon by means of dense deployment of sensor nodes (Vuran et al., 2004). Wireless sensor networks are usually composed of low-cost and low-power sensor nodes that make it possible to record observations with high spatial and temporal resolution.

Wireless sensor nodes are equipped with sensing, processing, and actuation capabilities (Liu et al., 2003) that are linked by wireless media to transform the required and partially processed data (Díaz-Ramírez et al., 2012). In recent years many applications have been proposed for WSNs. On the other hand, data for WSNs may observe by many different types of sensors such as seismic, low sampling rate magnetic, thermal, visual, infrared, acoustic, and radar (Akyildiz et al., 2002). Moreover, diverse variety of attributes can be observed that include the following: temperature, pressure, humidity, soil make up, vehicular movement, noise levels, lighting conditions, the presence or absence of certain kinds of objects, mechanical stress levels on attached objects, the current characteristics such as speed, direction, and size of an object and so on (Akyildiz et al., 2002; Estrin et al., 1999). Area monitoring, precision agriculture and horticulture, (near) real-time event detection, air pollution monitoring, landslide detection, local control of actuators, health, military, and security are only a few applications of currently used WSNs. However, data collection and monitoring are the main objectives through the WSNs.

Wireless sensor networks are installed with respect to the application and type of the required information, quality of the service or networks topology, communication protocols and routing, power management, network structure and hierarchical networks (Lewis, 2004) since aforementioned parameters may affect the quality of the derived data. WSN applications require efficient, accurate, and high spatial-temporal resolution data to address (near) real-time decision making and situation awareness. Quality of data provided by WSNs is highly critical while, provided raw data may be drawn in from of a low quality and reliability level expectedly, because of the sensors nature. The quality of sensed data may be affected by noise and fault, missing values, duplicated data or inconstant data (Zhang et al., 2010). Noise and fault, event, and malicious attacks can be considered as a main outlier source in WSNs (Zhang et al., 2010). Limited numbers of WSN resources, low quality ones, harsh deployment environments, memory, computational capacity, and computational bandwidth can cause unstable data (Zhang, 2010). Whereas sensor faults are not spatially dependent (Ding et al., 2005), events are spatially dependent in dense networks. Consequently, outlier detection procedure refines the measured data and leads to retrieve potentially useful information, called an event.

Events in different domains may interpret, differently. Event detection is counted as one of the main applications of the WSNs. While the purpose of the study is event detection, it is important to detect events in (near) real time. It can cause more challenge for online or (near) real-time event detection while, large quantities of spatial and temporal data cause difficulties with respect to the limited memory, computation, communication capabilities and constraint on sensor nodes sensing capability (Shih et al., 2008). The limited resources and capability, unreliable and imprecise sensor observations, and the dynamic changing of an environment, make WSN events detection challenging in the context of WSNs (Kapitanova and Son, 2010). Technology requirements of WSNs also draw special attention to the energy consumption while, all the accurate event detection methods could not be applied in WSNs because of energy constrains. Accuracy and energy efficiency are the main challenges in the event detection where indirect relationship exists among energy consumption and detection accuracy, generally. Whereas being energy efficient could increase the life time of sensors and somehow prevent some kinds of faults, accurate methods usually need to consume more energy. Generally, large amount of data, more communication and processing rate are prerequisite for accurate event detection. Hence, there is a trade-off between accuracy and energy efficiency to design the effective detection method. To evaluate the accuracy of the event detection method detection rate and false positive rate are assessed. Applied method also should be evaluated in terms of energy efficiency. Events maybe detected in WSNs individually in single nodes or collaboratively in multiple nodes. Spatial and temporal components of observations are essential where time and location of the events are of

interest (Kapitanova and Son, 2010). Inexpensive and low-power characteristics of the WSN devices encourage dense distribution of sensors that reminds the spatial correlation (Vuran and Akan, 2006). Moreover, Vuran and Akan (2006) mentioned consecutive temporal records are temporally correlated while, similar properties for spatially close observations are expected. So, by considering spatial and temporal properties of phenomena it is beneficial to utilize all the potentially useful available data by means of spatial and temporal models to obtain a better realization from phenomena. However, due to existence of faults or in temporary deployments lack of historical data is expected. Thus, performing the temporal correlation models is challenging.

While temporal and spatial properties of occurred events are important, temporal extent of an event should receive a more attentive consideration. In order to meet the event detection objective, it is important to identify an event and subsequently, track the spatial and temporal development of an event where event tracking requires frequent sampling and communication in order to sense event features.

Conventional event detectors, follows three type of scenario in statistic context:

- Detecting the events based on temporal correlation models.
- Detecting the events based on the spatial correlation models.
- Detecting the events by cooperating temporal models and spatial models.

Generally, sensing all the properties of an event is not possible, where most often an individual sensor node cannot detect the events properly. Thus, it is the emergence of using all the available potentially useful data. Events could not be distinguished sufficiently from outliers in the temporal models in the absence of multiple variables while, events are detected in individual sensor nodes in time and spatial data of neighbor nodes are ignored (Zhang et al., 2012). Moreover, Zhang (2010) mentioned, events could not be distinguished properly from long-term faults in the absence of enough information and neighboring nodes data. Spatial models detect the events with respect to the multiple nodes at a moment in time where, it ignores the time sequence data (Zhang et al., 2012). Considering the advantages and disadvantages of spatial correlation and temporal correlation in order to obtain accurate results spatial and temporal properties should be integrated. In addition, it is necessary to use multi-nodes approach in order to monitor and track the temporal evolution and geographic extent of events.

1.2 Research identification

The aim of this research is to detect outliers and events and identify the geographic extent of events and their evolution over time.

1.2.1 Research objectives

The main objective of this research is to develop a techniques to identify outliers and events with respect to the spatial and temporal properties of the WSN data in an energy efficient and accurate manner.

1.2.2 Specific objectives and research questions

In order to achieve the research objectives following specific objectives and questions should be addressed:

1. To define outliers and events.
 - 1.1. How can an outlier be defined?
 - 1.2. How can an event be defined?
 - 1.3. How can an outlier become an event?
2. To detect an event.
 - 2.1. How can temporal properties of WSN data be used to identify events?
 - 2.2. How can spatial properties of WSN data be used to identify events?
 - 2.3. How to integrate spatial and temporal properties of WSN to identify events?
3. To characterize an event.
 - 3.1. What is the temporal evolution of event?
 - 3.2. What is the spatial extend of event?
4. To evaluate the detected outliers and events.
 - 4.1. How can a reference data-set be identified?
 - 4.2. How accurately can outliers be detected?
 - 4.3. How accurately can events be detected?
 - 4.4. How efficient is the proposed event detection method in terms of energy?

1.3 Innovation

This research aimed at developing an energy efficient spatial-temporal outlier and event detection method. The main contribution of this research is to configure an energy friendly method for identifying the events based on the patterns by using the contextual information.

1.4 Thesis organization

The organization of the thesis is shown in Figure 1.1 which demonstrate the information flow of the main thesis topics, the contributions and relationship among the chapters. The rest of the thesis is organized as follows. Chapter 2 gives the theoretical background. Chapter 3 describes study area, dataset description, tools and data pre-processing. In chapter 4, the proposed methodologies are presented. Chapter 5 gives experimental results to demonstrate the proposed approaches. In chapter 6, discussions on methodology and results are provided. Thesis is concluded in chapter 7 by summarizing the key results, highlighting the lessons learned and direction for future work.

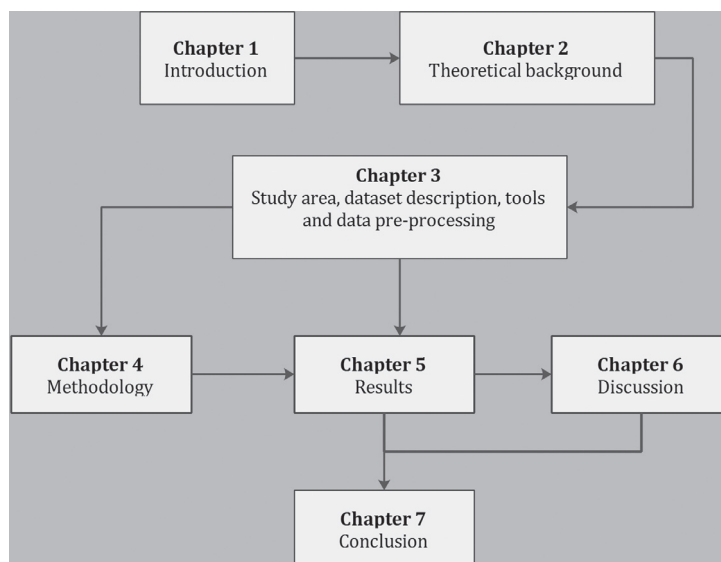


FIGURE 1.1: The organization of the thesis

Chapter 2

Theoretical Background

2.1 Outlier and event definitions

Application domain affects the definitions of outliers and events. For example, outliers can be defined for metrological observations based on standards that are provided by world meteorological organization (WMO). Zahumenský and SHMI (2004) proposed plausible value check and plausible rate of change for outlier defining. Zhang (2010) applied the plausible value check to fault detection by using pre-defined threshold. It should be consider that for defining events, availability of information such as nature of the variable, space and time specifications of the variable and events properties are important. In other words, a unique and constant threshold cannot be specified for all environmental observations and the threshold may vary by different times, days, weeks, months, locations and events. Moreover, even by specifying a threshold for a specific type of variable, in a specific time, events cannot be distinguished from outliers while most of the time event type is unknown in advance. Hence, it is not simply pragmatic and realistic, to have definitions for events and outliers based on a pre-defined threshold in the absence of prior knowledge on upcoming events (Zhang et al., 2009).

In order to distinguish events from outliers; data properties, application domain, and detection method should be taken into account. Generally, outliers are defined as data points which are very different from the rest of the data based on some measure (Aggarwal and Yu, 2001). Barnett and Lewis (1984) defined outliers as those observations that are inconsistent with other observations. Outliers may present any change in the state of the environment that is called event or include any source of faults.

Faults may include incidental absolute error, clustered absolute error, random error, long-term error (Zhang, 2010), dead band errors and systematic errors. Incidental absolute error or blunders are defined as a short period and unexpected errors while, clustered absolute errors have the same nature but occur in a long time period. Random errors are not within the threshold based on normal behavior while, long-term errors contain random errors properties but occur in a long time sequences (Zhang, 2010). On the other hand, systematic errors often occur because of sensors imperfection. Thus, by excluding faults from outliers, potentially useful information will be retrieved from outliers that illustrate a real change in the state of the environment. However, it is not usually straightforward to distinguish faults in terms of long-term error and systematic errors from events. Dealing with outliers is an area of interest while, they can be adjusted, reject or remain unaltered based on the type of outlier (Onoz and OGUZ, 2003). In the case of faults they can be ignored or replaced while, events that present a change in the state of environment should be remained for further investigations.

2.2 Spatial and temporal correlations in WSNs

Spatial and temporal properties of WSNs data, acknowledge applying of temporal and geostatistical analysis to obtain temporal and spatial movement and extent of events.

2.2.1 Spatial modeling by geostatistics

The mathematical difference of any pair of observations which is calculated by half of the squared difference between both values presents semivariance. The usual formula for computing the the sample Variogram is:

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{s=1}^{n(h)} (y_s - y_{s+h})^2 \quad (2.1)$$

where: s is a vector of spatial coordinates, h is the lag distance representing separation between two spatial locations, y_s is the variable under consideration, y_{s+h} is the same value of a separation of h , s presents the different spatial locations of the data, and $n(h)$ is the number of point pairs (y_s, y_{s+h}) separated by h .

Variogram can be described based on the follow characteristics by Figure 2.1:

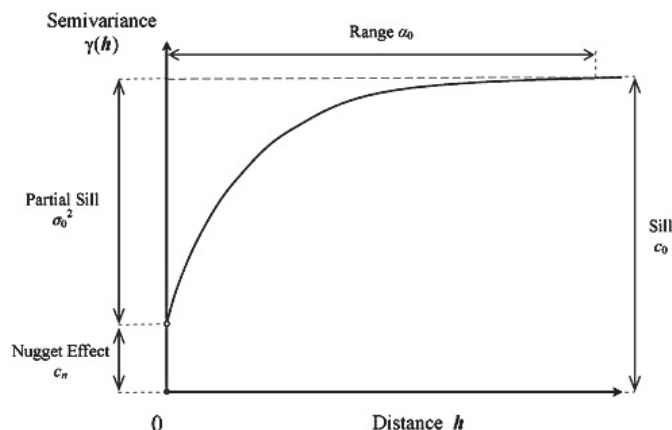


FIGURE 2.1: A theoretical Gaussian model variogram

- The Sill: Correspond to the maximum semivariance value, which represents the values at the locations that are spatially independent. The difference between the sill and the nugget (described below) is known as the partial sill.
- The Range: The range presents the corresponding separation distance at which the sill is reached.
- The Nugget effect: The nugget effect represents a discontinuity of the semivariogram that can be present at the origin. As the distance goes to zero, there tends to be a nugget effect due to measurement errors and non-spatial and micro-scale variation.

2.2.1.1 Prediction

Fitted model to the variogram and coordinates of the neighboring locations and corresponding observations are used for predicting an unknown observation at a known location. An optimal geostatistical interpolation technique that provides the best linear unbiased predictor (BLUP) for every location is called kriging. Kriging is used to predict the unknown observations as a linear weighted combination of observations collected at its neighbor location.

2.2.2 COSMO forecasting models

In the absence of sufficient historical data to build temporal correlation model to build the temporal correlation model climatic forecasts are utilized. In this research, instead of building the temporal correlation model by WSN data, climatic forecasts are utilized as a result of

temporal correlation model to perform temporal outlier detection. Forecasts are provided by MeteoSwiss (2013) for the study area.

Consortium for small-scale modeling (COSMO) operates and further develops a high-precision numerical weather prediction system, to automatically generate regional and local forecasts products in complex topography (Baldauf et al., 2011). A detailed image of the future state of the atmosphere is computed, from the low stratosphere to the surface, including the evolution of the snow cover, the lake temperature and the soil characteristics. COSMO-7 produces forecasts up to three days in advance on a domain covering central and western Europe. Figure 2.2 shows the model suite of MeteoSwiss: the european center for medium range weather forecast (ECMWF) operates a global model describing the synoptic scales. MeteoSwiss operates the regional scale COSMO-7 and the local scale COSMO-2 models.

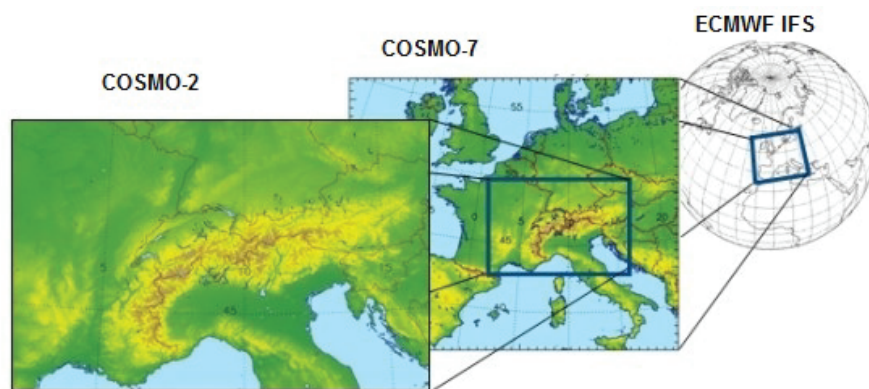


FIGURE 2.2: The three nested numerical weather prediction models of the COSMO system (MeteoSwiss, 2013)

A numerical weather prediction model is based on the physical laws describing the atmospheric and surface processes (e.g. conservation of energy, transformation of phase, black body radiation), suitable initial and boundary conditions, and a numerical method to solve in time the resulting system of complex mathematical equations (Baldauf et al., 2011). These equations describe a variety of atmospheric processes on different temporal and spatial scales, for example the development of low pressure systems, snow fall and summer convection. The calculations are performed on a three-dimensional grid, where the vertical spacing of the computational levels is inhomogeneous in order to better resolve the low level atmospheric conditions (Baldauf et al., 2011).

Suitable initial conditions for starting a forecast are produced by combining actual atmospheric observations, previous model guidance, and climatic information in a way compatible with the

model equations; this is the so-called data assimilation process. In a typical 24 hours assimilation period, COSMO-7 ingests (Baldauf et al., 2011):

- about 120 vertical soundings,
- about 8000 aircraft observations,
- about 28000 surface observations, and
- about 1000 wind profiler measurements.

COSMO-2 and COSMO-7 models provide operational products as following:

COSMO-2, with a mesh size of 2×2 km, COSMO-7, with a mesh size of 7×7 km.

2.2.2.1 Localizing the forecast values

To localize the forecast values, algorithm searches grid points in the vicinity of the surface station, consequently optimization is performed in terms of horizontal distance and vertical height difference (Kaufmann, 2008).

2.2.2.2 Search radius depending on model surface at station

Localization is performed, by determining the land surface type (land or water) of the model at the station location, firstly. Then grid cell, which the station is located in, can be determined by rounding the grid coordinates of the station to the nearest whole number. Depending on the surface type of the respective model grid cell, the search radius for model grid points around the station is defined as (Kaufmann, 2008):

$$r_{sreach} = r_{land} = 1.415 \text{ for land surface,}$$

$$r_{sreach} = r_{water} = 2 \text{ for water surface, in model grid units.}$$

The value of the radius is chosen to keep the method in close accordance with the previously used method but to allow a wider search area over a flat water surface.

2.2.2.3 Search area and preference of land points

All grid points within a horizontal distance of research of the exact station location are evaluated. The number of evaluated grid points depends on the location of the station relative to the model grid. If at least one model grid point within the search radius has a land surface, all water points are excluded from the selection.

2.2.2.4 Optimization of horizontal and vertical distance

Horizontal distance $d_{hor} = \sqrt{\Delta x^2 + \Delta y^2}$ and the vertical height difference $d_{ver} = z$ of the station to all grid points are calculated, in the same geometrical length unit, e.g. in meter. Aforementioned two elements are combined to an optimization distance according to Kaufmann (2008):

$$d_{opt} = d_{hor} + |d_{vert}| \cdot f_{ve}$$

with the vertical emphasis factor $f_{ve} = 500$. Smallest optimization distance grid point is selected, consequently. This grid point shall be the model grid point associated with the station.

2.3 Related work

In this section, conventional statistics-based outlier and event detection techniques are presented. In order to provide an easier overview, they are categorized to the spatial, temporal, and spatial-temporal event detection techniques in the context of WSNs.

2.3.1 Spatial outlier and event detection methods

Wu et al. (2007) proposed a localized algorithm which applied implicitly spatial correlation to distinguish events from outliers at an event boundary. Their proposed method identifies outliers based on differences of observed value of each station and the median reading of adjacent sensors. Subsequently, each node collects all differences from its neighborhood and standardizes them. Finally, outlier is occurred when the absolute value of its standardized difference is sufficiently larger than a preselected threshold. They also proposed a scheme to adaptively adjust the thresholds in the base station. Yang et al. (2008) proposed an unsupervised (one-class) centered quarter-sphere support vector machine method and took advantage of spatial correlations that exist among sensor data of adjacent nodes to reduce the false alarm rate in real-time. Wang and

Yu (2005) proposed a method for event region detection based on the Markov Random Field (MRF) model and a sensor fusion technique. Event regions were identified where MRF model was used to model the spatial correlation, and the sensor fusion technique considered the concept of sensor reliability. Jin and Nittel (2006) proposed a distributed algorithm that utilized implicitly spatial correlation among nodes to event and event Boundary detection. Zhang et al. (2012) proposed a spatial outlier detection method in which nodes transmit their observations to the adjacent nodes in every sensing instant in time and while a node detects the outlier, it sends notification to the neighboring nodes. They detected events in the presence of confirmation from adjacent nodes. This technique was not energy efficient and could not address the online applications while, large amount of data transition and communication exist. Yu et al. (2012) proposed event detection method based on spatial correlation, to perform prediction to find the event features, correct the faulty sensor nodes and detect the event region.

2.3.1.1 Spatially based event definitions for outliers and events

Outliers are spatially independent, whereas normal observations of the same phenomenon are geographically correlated (Krishnamachari and Iyengar, 2004). Spatial outliers are those observations that difference between predicted value/pattern and observed value/pattern exceed pre-defined threshold or lies outside of the confidence interval. Predicted values can be obtained from the spatial correlation model. Generally, it is assumed that events are spatially correlated. Events are occurred when multiple number of outliers are happened in correlated adjacent nodes at a specific time instance. While outliers can be occurred in each of the attributes independently, co-occurrence of outliers in multiple attributes in single node by using spatial predictions can present events.

2.3.2 Temporal outlier and event detection methods

Subramaniam et al. (2006) utilized proposed an outlier detection method that utilized temporal correlation. They labeled observation as an outlier where each node identified local outliers while the observation deviated significantly from the result of temporal model. Zhang et al. (2012) proposed an online temporal outlier detection method by using the time series analysis. Their method utilized the predictions of autoregressive model to identify outliers. This method identify outliers where an observation lays outside the confidence of its predicted value. Ogbagabriel (2012) proposed a temporal outlier detection method based on maximum variability check. This method identified outliers where observations jump over a specific time period, unrealistically.

Ogbagabriel (2012) proposed a method to identify temporal outliers by using moving average and moving median.

2.3.2.1 Temporal based definitions for outliers and events

Temporal outliers may be those observations that jump in time at successive records in a single node. Subsequently, in the presence of multiple variables, co-occurrence of outliers in single node at a multiple attribute in a specific time instance can illustrate events. It should be noted that outliers can be occurred in each of the attributes independently. Moreover, events can be determined based on the length of outlier sequence. In other words, based on duration of the outliers, events can be identified (Zhang et al., 2012) in multiple nodes and it is highly effected by application requirements and sampling rate. Outliers can be defined as those observations that deviate considerably from predicted value/pattern with respect to the pre-defined threshold or placing outside of the confidence interval. Temporal correlation models are used to predict a value. Hence, to identify events co-occurrence of multivariate outliers and the length of outlier sequence are applicable.

2.3.3 Spatial-temporal outlier and event detection methods

Bettencourt et al. (2007) developed a distributed method to identify outliers and events in ecological application of WSN. This method addressed the non-stationarity by learning the statistical distribution of difference between its own measurements and each of the adjacent nodes, as well as its current and previous measurements. Outliers were identified where measurements were less than a user-defined threshold. Consequently, events were detected since outliers were different temporally from previous observations and spatial correlation existed by other outliers. Wang et al. (2008) proposed combination of dimensionality analysis and a Bayesian networks as a means for unsupervised learning and anomaly (event) detection in gas monitoring sensor networks for underground coal mines. Wang and Wu (2010) utilized spatial-temporal correlation based outlier detection in which distributed algorithm utilized spatial and temporal correlation between sensor nodes to detect outliers. In order to identify outliers, similarity of two successive sliding window observation were checked and in the presence of considerable jump it would continue by checking whether it is faulty data or not. Then, they used spatial correlation of last sliding window among different nodes to check which nodes were interrelated and by using pre-defined threshold spatial correlated nodes can be determine. Moreover, if the current sliding window of a node and correlated neighbors own sliding window presented similarity then a

positive vote will be given otherwise a negative vote and non-correlated nodes do not affect the procedure. Finally, negative sum of votes can illustrate the faulty observations of corresponding sliding window.

Shahid and Naqvi (2011) proposed a quarter-sphere SVM based outlier and event detection. Their method utilized temporal correlation along with the attribute correlations at each node. One class Quarter-Sphere SVM based on temporal and attribute correlations were used for outlier detection. They identified outliers based on distance criteria and events were defined with respect to the decision criteria which counts the number of sensor arrays (group of sensors which measure different attributes) whose distances are greater than their corresponding radii. They utilized spatial correlation to identify events, implicitly. Events were defined as those observations where an outlier on at least half of the sensor nodes occurred. Dereszynski and Dietterich (2011) proposed a model that adapted to each deployment site by learning a Bayesian network structure that captured spatial relationships between sensors, and it extended the structure to a dynamic Bayesian network to incorporate temporal correlations. This model had a capability to label faulty observations and predict the true values of the missing or corrupt readings. Their process model incorporated spatial component that represents the relationships among all sensors within a deployment for a given time slice and temporal component that captures the transition dynamics from one time period to the next. Outliers were defined by inferring the most likely state of the sensors given the current temperature observations and those of the immediate past.

Shahid et al. (2012) proposed SVM based method by cooperating Quarter- Sphere (QS) formulation of one-class SVM. They applied QS-SVM formulation based on attribute and spatial-temporal. They modified QS-SVM by utilizing attribute correlation while, conventional QS-SVM methods just used spatial-temporal correlation. So, they improved the results by using attribute correlation as well as spatial-temporal correlation and results showed that spatial-temporal and attribute QS-SVM increase the detection rate considerably over the spatial-temporal QS-SVM. They used advantages of QS-SVM in their work to provide online, adaptive, and distributed algorithm. They performed the event detection by detecting the radii of QS based on spatial-temporal and attribute correlation. They presented applying attribute correlation in the presence of spatial-temporal correlation in QS-SVM to improve outlier detection rate and reduce false positive rate over the spatial-temporal QS-SVM. Albanese et al. (2012) proposed a method, called rough outlier set extraction from spatial-temporal data that relies on a rough set theoretic representation of the outlier set using the rough set approximations, i.e. lower and upper approximations. Moreover, they also applied a new set, named Kernel Set,

that is a subset of the original data-set, which is able to describe the original data-set both in terms of data structure and of obtained results.

Zhang et al. (2012) proposed three spatial-temporal correlation based outlier detection. In first method, they integrated spatial and temporal correlation while, each node identified temporal outliers and then outliers were distinguished from events by using spatial outliers results and prediction was accomplished with respect to the actual measurements. They proposed a second method based on spatial predicted-data-based outlier detection which, used parameters of autoregressive model and neighboring data to prediction. Thirdly, they proposed spatial and temporal integrated method to outlier and event detection in which, weights of nodes and temporal correlation parameters were applied separately.

2.3.3.1 Spatial-temporal based definitions for outliers and events

For spatial-temporal outlier and event detection two scenarios can be supposed: first using spatial and temporal models separately and second using integrated spatial-temporal model. Firstly, spatial and temporal separate models can detect outliers by using temporal models. Hence, outliers are defined with respect to considerable deviation from predicted value with respect to the pre-defined threshold or confidence interval and events are presented by co-occurrence of multivariate outliers and on the other hand length of the outlier sequence is applicable for event detection. Secondly, there are methods that utilize spatial and temporal models as an integrated model. They also identify outliers in successive temporal observations then events can be detected based on length of outlier sequence. So, outliers are those observations that differ from predicted value by spatial-temporal model and events are those observations that occurred in successive multiple time sequence. Moreover, co-occurrence of spatial-temporal outliers in multiple correlated nodes can also present an existing of events.

As a conclusion, spatial event detection methods are incomplete simply while, temporal observations are ignored (Zhang et al., 2012) and due to inexpensive and imprecise nature of the WSNs multiple faulty nodes can happen, frequently which may be labeled as event. Zhang (2010) mentioned events that are detected by temporal correlation cannot be reliable while neighboring data are not considered and events are detected locally. In addition, because of not existing the spatial information a single node cannot distinguish long-term faults from events. Temporal event detection methods cannot be applied in the absence of correlated multiple attributes which are not available, generally. In order to overcome disadvantages and drawbacks of spatial correlation and temporal correlation event detection methods, an integrated approach should

be applied to utilize both spatial and temporal properties of WSN data to event detection in WSNs.

Chapter 3

Study area, Dataset description, Tools and Data pre-processing

3.1 Study area

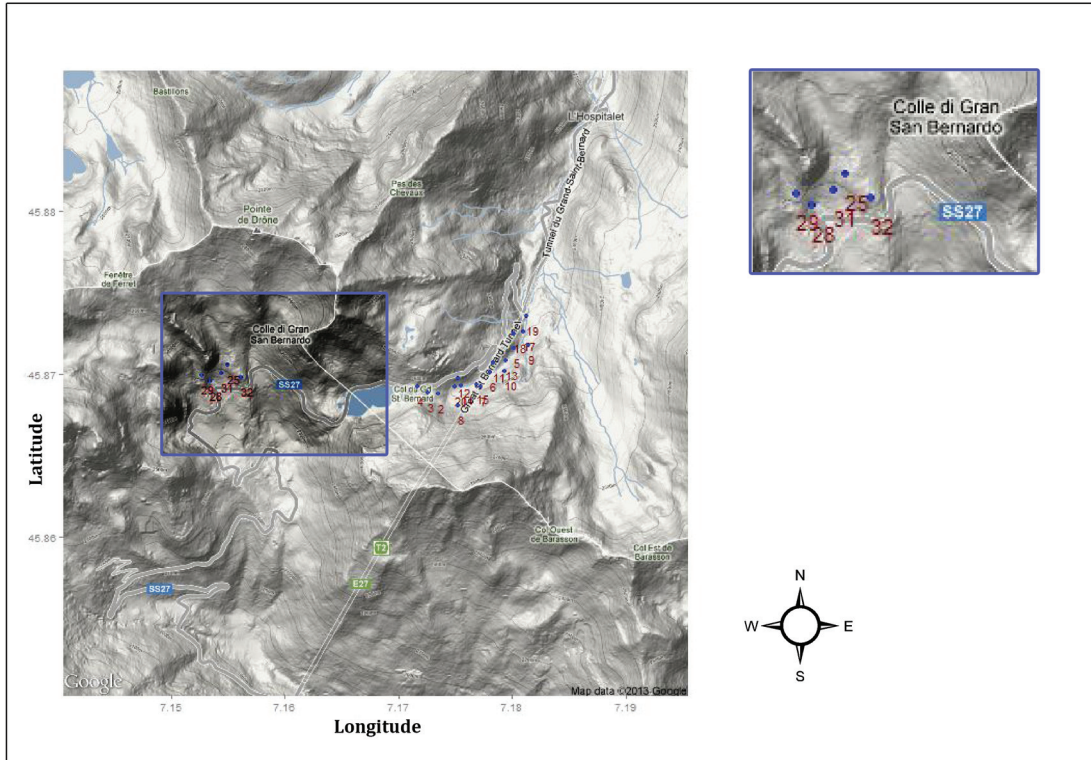
Grand-St.-Bernard wireless sensor network is deployed at the Grand-St-Bernard pass at 2400 m between Switzerland and Italy. Most of the stations are located in Switzerland and a few are installed behind the Italian border that are presented by Figure 3.1 .

3.2 Wireless sensor network dataset

The Grand-St.-Bernard data-set is investigated in this research (Figure 3.5). The dataset is provided by interdisciplinary project called SensoreScope in 2007. The project aims to develop a new generation of measurement system based on a wireless sensor network with built-in capacity to produce high temporal and spatial density measures, push the limits in real-time environmental measurement by deploying sensing stations in the Swiss Alps and development of a complete communication stack well-fitted to harsh deployment sites featuring multi-hop routing and synchronization among stations, as well as an advanced energy management of the radio chip (Sensorscope, 2007).

The elevation range in the study area is from 2300 meters to 2500 meters. The available dataset was observed from 13th September 2007 to 26th October 2007 and recorded at coordinated universal time (UTC) time zone for 45 days (Ingelrest et al., 2010). The frequency of the sampling for the deployment is two minutes. The data-set is provided by 23 sensor nodes.

a) The distribution of the nodes in the study site



b) The coordinates of nodes according to the Geographic coordinate system

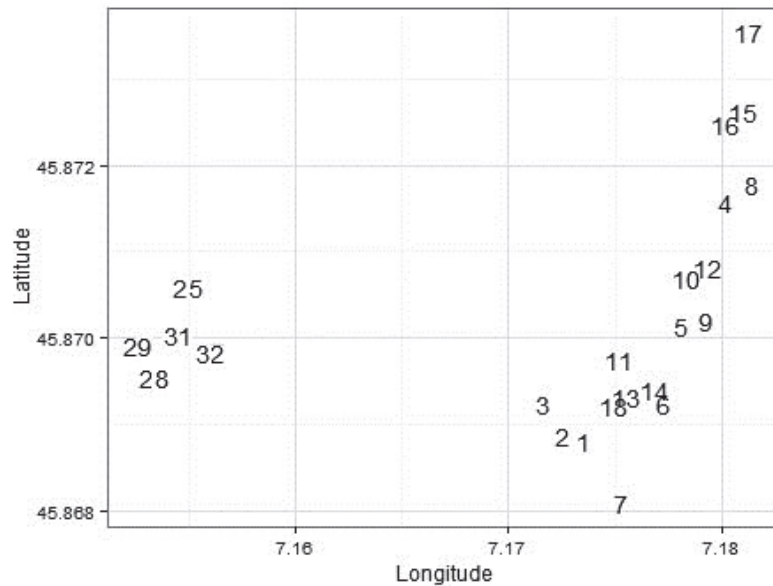


FIGURE 3.1: The Grand St. Bernard wireless sensor network deployment

Eighteen are deployed in Switzerland and rest of them are installed across the Italian border. These sensing stations periodically sample and transmit their readings through the wireless network to a collection points that are called base-stations. The nodes are deployed in two clusters. The big cluster consists of 18 nodes while, the small cluster includes five nodes.

The data-set includes seven external sensors, which makes the station capable of measuring nine different data inputs: ambient temperature and humidity, infrared surface temperature, solar radiation, wind speed and direction, precipitation, soil moisture, and soil pressure (Sensorscope, 2007). Data are collected under specific range and accuracy as presented by Table 3.1.

TABLE 3.1: Sensors specifications

Measurement Type	Range	Accuracy
Air Temperature	-20°C–60°C	±0.3°C
Humidity	0–100%	±2%
Surface Temperature	-20°C–70°C	±0.6°C

The proposed methodology was experimented on ambient temperature with respect to the small cluster consisting of sensor nodes 25, 28, 29, 31 and 32 that are highlighted in Figure 3.1(a). Observations were recorded at the same point in time, for the period 00:00~24:00 on two successive days, 29 and 30 September 2007 and presented in Figure 3.4.

3.3 MeteoSwiss dataset

To perform temporal outlier detection climatic observations and forecasts are utilized in this study that are provided by MeteoSwiss (2013).

3.3.1 MeteoSwiss observations

Hourly mean data of air temperature 2 meter above ground is provided by MeteoSwiss (2013) from 10th September 2007 to 28th October 2007. MeteoSwiss observations (MOs) are obtained from a station located in 'Col du Grand St-Bernard' in the 2472 elevation (Figure 3.5). Observation were recorded in UTC time zone. Hourly mean values (HH = (HH-1):41 - HH:40) were calculated for MeteoSwiss observations (MOs) with six values of temperature, measured in 10-minute intervals: (HH-1):50, HH:00, HH:10, HH:20, HH:30, HH:40 (MeteoSwiss, 2013).

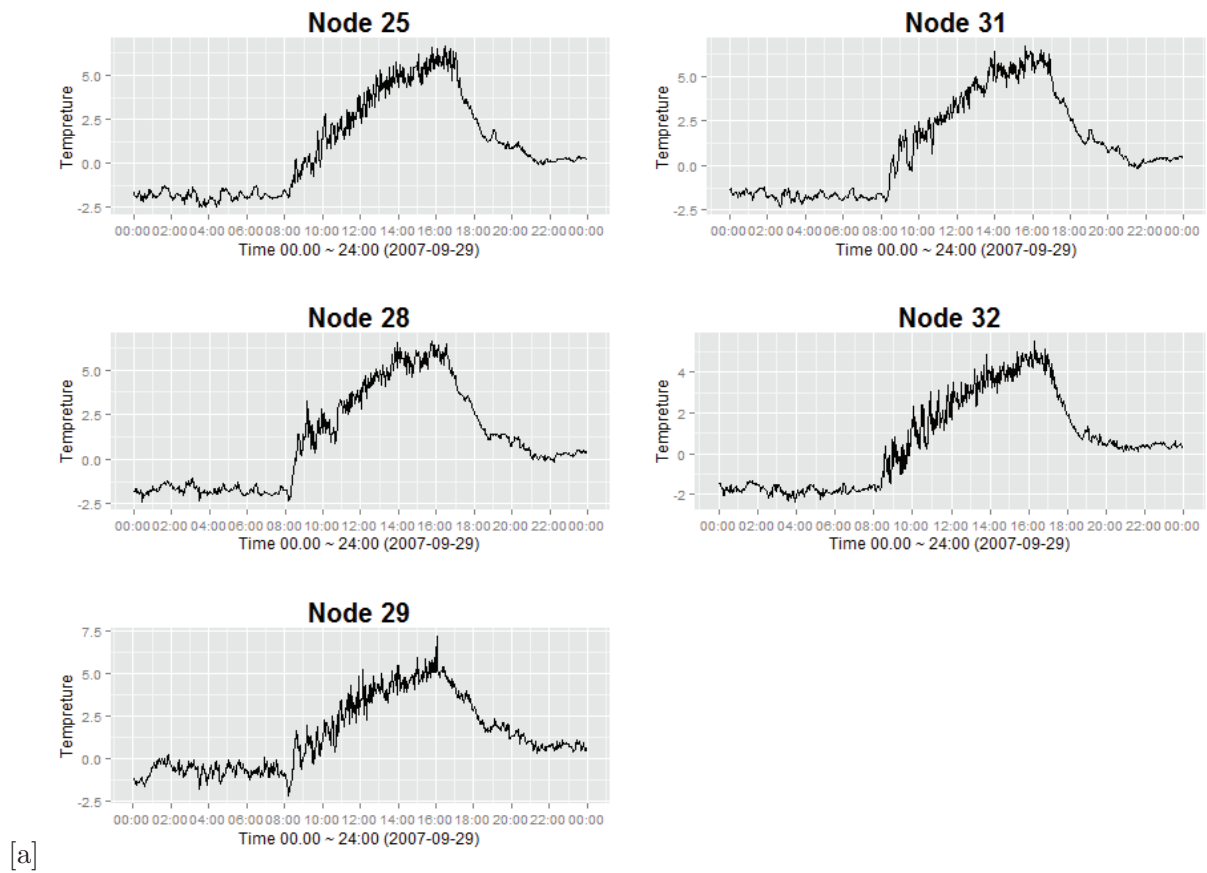


FIGURE 3.2: The illustration of ambient temperature on 2007-09-29

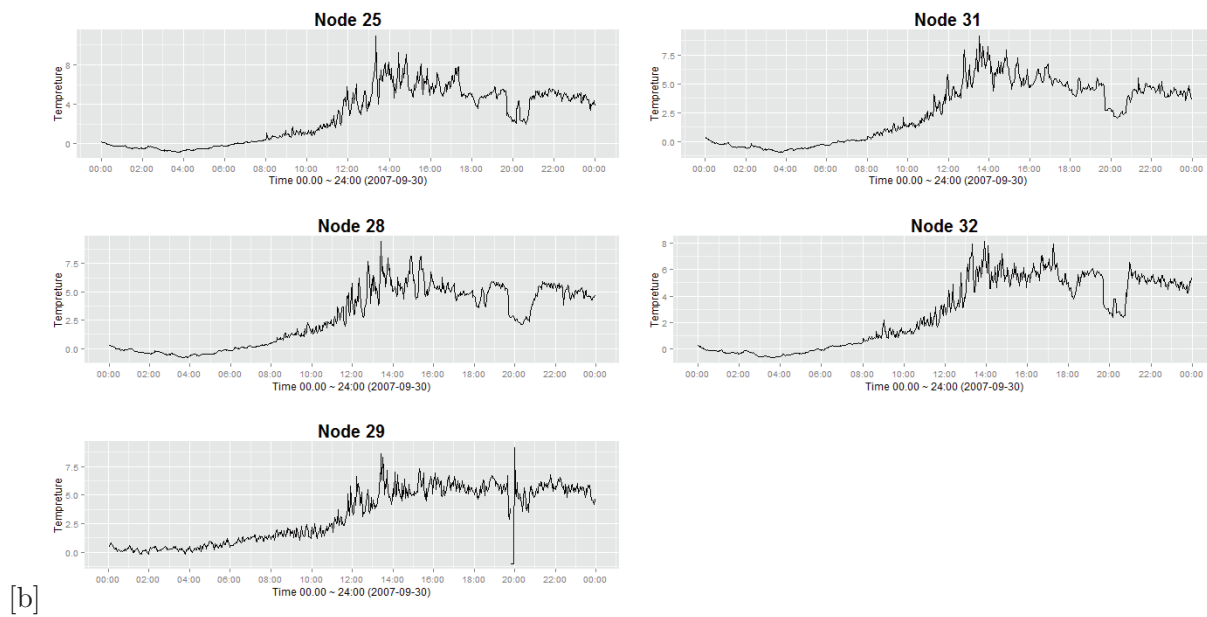


FIGURE 3.3: The illustration of ambient temperature on 2007-09-30

FIGURE 3.4: Illustration of ambient temperature on 2007-09-30 and 2007-09-29

3.3.2 MeteoSwiss forecasts

Forecast data are provided by MeteoSwiss (2013) that are hourly values 72 hours in advance. Model that is explored in section 2.2.2 resulted the hourly temperature forecasts. COSMO-7 forecast data up to 72 hours for the Grand-St.-Bernard and the four surrounding grid points (Figure 3.5) are retrieved from the operational $7\text{km} \times 7\text{km}$ versions of the COSMO at the study area. The grid point M that is presented in Figure 3.5 is the most representative grid point for a certain weather station Mo by looking to distance and height difference between model and real orography as explained in detailed in Section 2.2.2.1. COSMO-2 could be led to more local forecasts but due to this fact that COSMO-2 was not operational in 2007 we have to use COSMO-7 forecasts. MeteoSwiss forecasts (MFs) are provided with respect to the UTC time zone. Forecasts are instantaneous values but the model time step is 60s only and the real time resolution thus might be about five minutes.

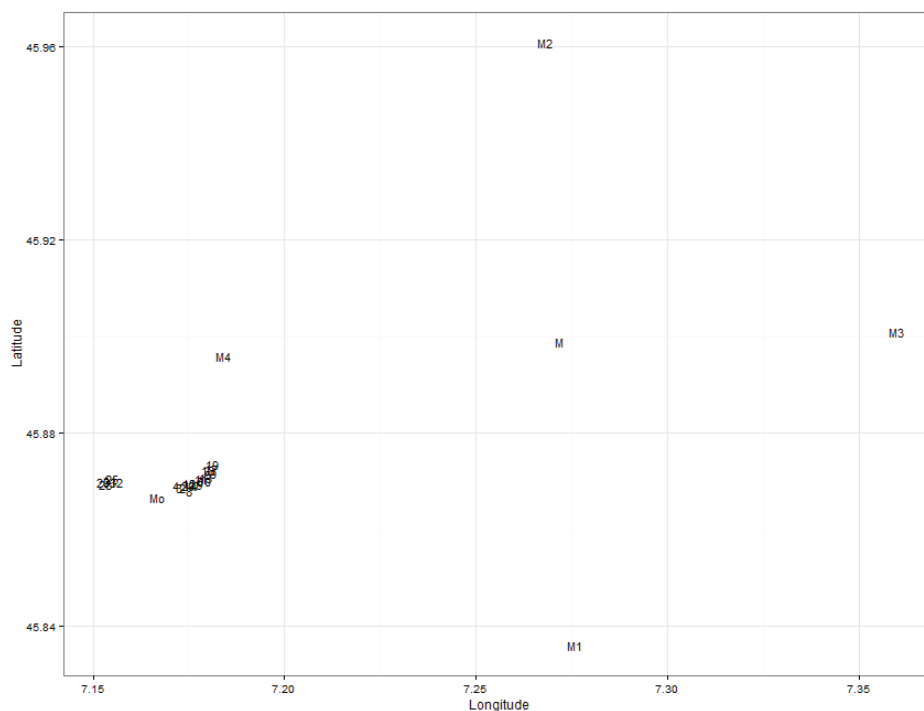


FIGURE 3.5: Locations of applied datasets in the research, where M , $M1$, $M2$, $M3$ and $M4$ present the MeteoSwiss forecast grids and Mo illustrates the MeteoSwiss observation station and two black clusters show the WSN small and big clusters in geographic coordinate system

3.4 Labeled dataset

In the absence of reference dataset, evaluation of outlier and event detection techniques is challenging. To assess the detection accuracy, labeled data from Zhang (2010) is used. The

dataset was provided by using three techniques: running average-based, Mahalanobis distance-based and density based. Moreover, aforementioned techniques resulted outliers for spatial, temporal and spatial-temporal domains on 2007–09–30 from 06:00 ~ 14:00 for nodes of the small cluster.

- Zhang (2010) performed the running average-based labeling technique by using a smoothing window and calculating the mean value for a fixed sample size. Subsequently, outliers are defined by taking the absolute value of the difference between the measurements and the values calculated by applying the running average filter. So, outliers are those observations that the absolute value of the difference between the measurements and the calculated values exceeds the certain threshold.
- Mahalanobis distance-based labeling technique was utilized by Zhang (2010) to identify outliers based on the measure of full dimensional Mahalanobis distance between a point and its nearest neighbor in the dataset. Hence, outliers are those observations whose Mahalanobis distance is larger than a predefined threshold.
- Zhang (2010) applied the density-based labeling technique to identify outliers by using diverse clusters in the dataset. Consequently, outliers are those observations if they locate in an area of the grid whose density is lower than a fixed percentage of the density values.

3.5 Tools

In order to accomplish the implementations in this research, two software are applied, simultaneously. In order to manage the WSN data PostgreSQL database system is applied which is a powerful, open source object-relational database system (PostgreSQL, 2013). While most of statistical analysis in this study are performed in R software so, RPostgreSQL (Conway et al., 2007) package of R is used in order to connect database to R. RPostgreSQL package provides a database interface compliant driver for R to access PostgreSQL database systems. In other words, all the commands and stored data of PostgreSQL can be accessed easily in R by using RPostgreSQL package.

In R software in order to perform analysis and presentation number of packages are applied. In order to perform spatial and geostatistical modeling, prediction and variogram map plotting gstat package is applied (Pebesma, 2004). Classes and methods for spatial data are provided by sp package (Pebesma and Bivand, 2005). In order to perform statistical analysis including traditional, likelihood-based and Bayesian methods geoR package is applied (Ribeiro Jr and Diggle,

2001). RPostgreSQL package is a tool for interaction of PostgreSQL and R software (Conway et al., 2007). The ggplot2 package is applied as an implementation of the grammar of graphics in R (Wickham, 2009). The scales package provides methods for automatically determining breaks and labels for axes and legends (Wickham, 2012). The reshape2 package provides flexibly restructure and aggregate data using just two functions: melt and cast (Wickham, 2007). Geographical maps are drawn by maps package (Brownrigg, 2012). The package of ggmap is applied for spatial visualization with Google Maps and OpenStreetMap (Kahle and Wickham, 2012). In order to convert latitude-longitude into projected coordinates mapproj package is utilized (McIlroy, 2003).

3.6 Data pre-processing

Data cleaning usually constitutes the initial step in wireless sensor networks (WSNs). In WSNs because of inexpensive sensors and harsh deployment, data need to be refined. Besides, WSN data are considered as streaming data since it is the emergence of performing data management because of large volume of data. Certain properties of a WSN dataset, such as dead band error and gross error can corrupt the analysis. Thus, cleaning the data is essential. In this study with respect to the nature of the WSNs data, plausible value check and minimum variability are applied in order to eliminate gross errors and dead band caused errors.

3.6.1 Database management system

To manage the WSN data, PostgreSQL database system is applied which is a powerful, open source object-relational database system (PostgreSQL, 2013). While most of statistical analysis in this study are performed in R software so, the RPostgreSQL (Conway et al., 2007) R package was used in order to connect database to R. RPostgreSQL provides a database interface compliant driver for R. In other words, all the commands and stored data of PostgreSQL can be accessed easily in R using RPostgreSQL.

3.6.2 Data cleaning: Gross error identification

Plausible value check (Zahumenský and SHMI, 2004) is used to identify gross errors for climatic data based on world meteorological organization (WMO) guidelines. Maximum and minimum limitations of observation are checked by verifying if the observed values are within the acceptable range limits. Acceptable range limits are defined based on sensing capability range of

the *Grand – St. – Bernard* sensors which are presented in Table 3.1 and with respect to the geographic location and season. Subsequently, those observations that are exceeded the limit are detected as faults and are not used for further analysis.

For this research gross errors are defined based on the historical data range of temperature in the Swiss Alps for the months of September and October of the last 10 years as those observations that lay outside the minimum of -5°C and maximum of 25°C (MeteoSwiss, 2013) and those observations that exceed the aforementioned threshold will not be used for further analysis (Figure 3.6). Based on the acceptable range number of observations are detected as fault that are presented in Table 3.2.

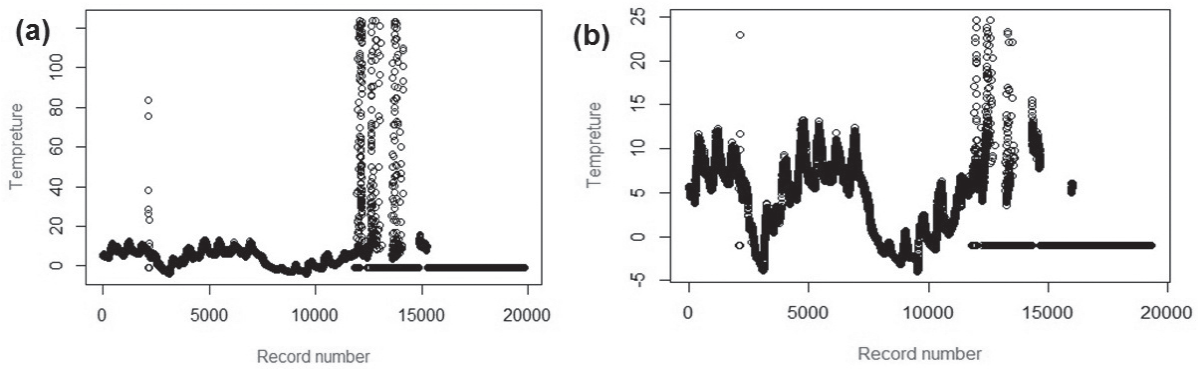


FIGURE 3.6: An example of performing the plausible value check for ambient temperature, they are laid outside the $(-5^{\circ}\text{C}, 25^{\circ}\text{C})$ threshold (a) before performing the plausible value check (b) after performing the plausible value check

TABLE 3.2: Gross errors found using plausible value check for ambient temperature of nodes [25–28–29–31–32]

Type	Node name				
	25	28	29	31	32
A	20276	20302	20267	20263	20257
B	1376	1321	83	1329	1502
C	0	0	0	0	0

A: Total number of observations

B: Number of faults for the whole data set

C: Number of faults on 2007–09–29 and 2007–09–30

3.6.3 Data cleaning: Identification of dead band error

A time consistency check (also called plausible rate of change) is used to verify the temporal consistency of instantaneous data among current and preceding values (Zahumenský and SHMI, 2004). If the difference between the current value and previous observation exceed a specific limit then the current value labeled as outlier. In order to identify dead band error, the minimum variability check was performed. To perform the minimum variability check, a cluster

of successive observations are checked if they do not vary over the pre-defined time period by more than a specific limit, all the corresponding observations are considered as fault. To define a threshold for minimum variability check elements such as accuracy of the sensor, geographic location and season are considered.

To define a threshold for minimum variability check elements such as accuracy of the sensor (Table 3.1), geographic location and season are considered. Subsequently, for ambient temperature based on WMO guidelines (Zahumenský and SHMI, 2004), minimum variability threshold is defined $\pm 0.5^\circ$ to cover a period of 60 minutes. Table 3.3 illustrates the quantity of the faults that are occurred because of dead band.

TABLE 3.3: Dead band caused errors found using plausible rate of change for ambient temperature on nodes [25–28–29–31–32]

Type	Node name				
	25	28	29	31	32
A	20276	20302	20267	20263	20257
B	214	65	4316	335	153
C	0	0	0	0	0

A: Total number of observations

B: Number of faults for the whole data set

C: Number of faults on 2007–09–29 and 2007–09–30

Consequently, six number of nodes on 2007–09–29 and 2007–09–30 are ignored from further analysis. Nodes 6, 9, 11, 15, 18 and 20 are not utilized in process while, they include considerable number of missed values, gross errors and dead band errors.

Chapter 4

Methodology

The methodologies of the research for outlier and event detection are illustrated by Figure 4.1.

4.1 Temporal outlier detection

To perform a successful temporal modeling long-term records should contain seasonal and diurnal effects. Seasonality and diurnally effects are necessary as the predicted variables illustrate seasonal and daily effects, specifically in outlier and event detection domains. In this research, the main focus is on ambient temperature as a climatic variable. Note that the dataset, collected over a 45 day period, includes considerable numbers of missing data, gross errors and dead band caused errors is presented in Table 3.2 and Table 3.3. Moreover, the problem of insufficient historical data also exists in the temporary sensor deployments, whose durations can range from a single week to several months. Subsequently, the amount of data decreased due to the existence of missing values, gross errors and dead band errors, which have been removed from the dataset. However, replacing the removed and missing values is beyond the scope of this research. Furthermore, the dataset does not include seasonal effects for accurate prediction of upcoming environmental variable.

While temporal correlation models typically produce accurate detections (Zhang et al., 2012), given their high detection and low communication rates, they cannot be utilized in this study, given the aforementioned restrain. Consequently, this research explores the use of information obtained from climate stations which is widely available and can be accessed from precise sensors in the presence of large amounts of historical data. Climatic weather stations usually include two general types of products, observations and forecasts. Hence, this research examines their

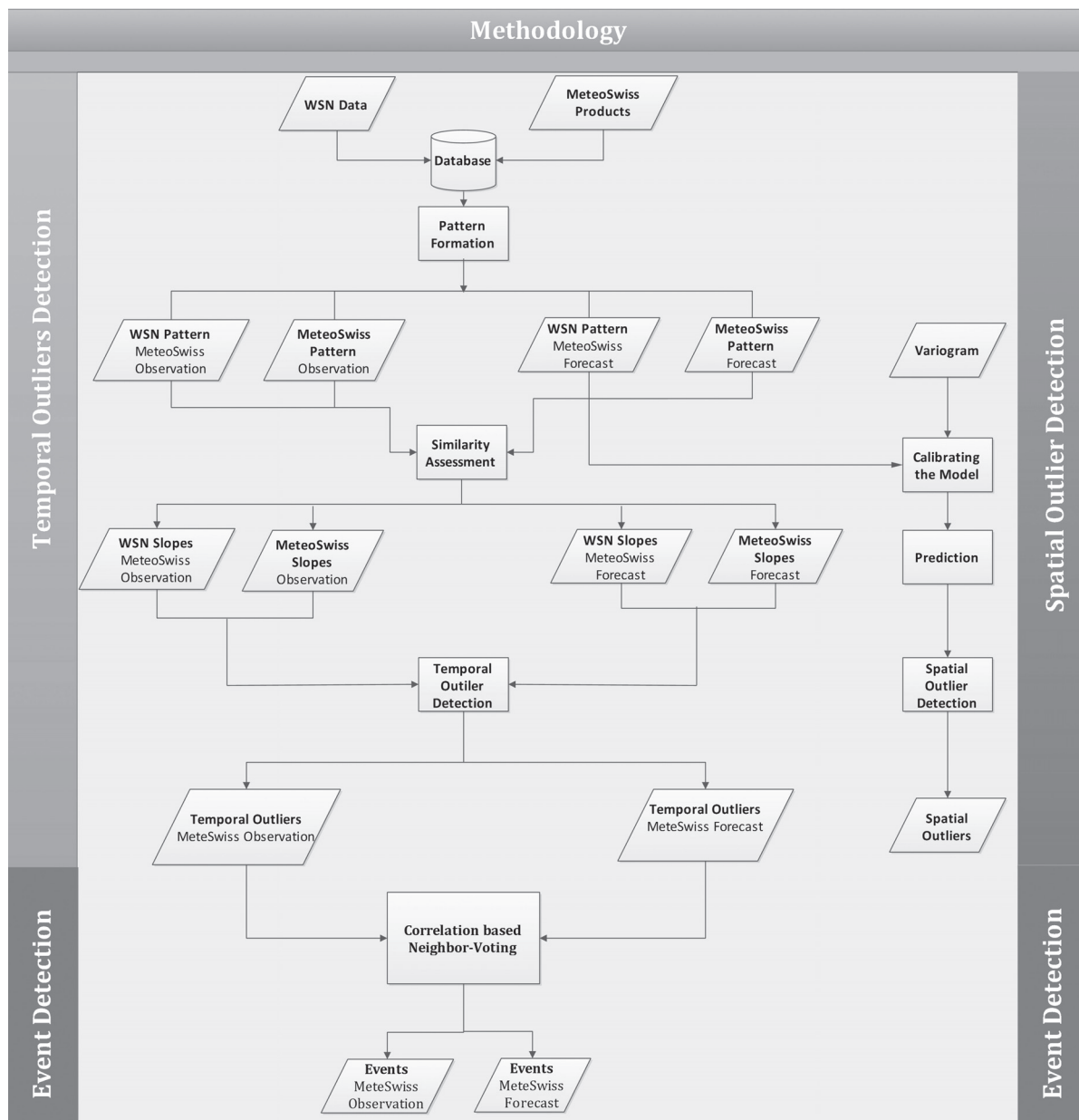


FIGURE 4.1: The flow chart of applied methods for outlier and event detection

applicability in the WSN context. Real observations and forecasts were provided by MeteoSwiss (2013) and are applied in this research. Thus, MeteoSwiss observations and forecasts as climatic products, are considered the results of a temporal correlation model, used to identify temporal outliers. Consequently, temporal outliers are those observations of WSN that differ from MeteoSwiss observations or forecasts.

WSNs may use either climatic forecast values in advance to identify temporal outliers or utilize observation values of climatic stations to build temporal models. WSNs usually deployed in harsh, uncontrolled or even hostile environments (Ma et al., 2013) with complex topography, thus it is not realistic to expect similar values for observations or forecasts of climatic

stations with respect to the wireless sensor measurements. The aforementioned argumentation consequently justify the comparison of patterns rather than absolute values. As presented by Table 4.2, WSNs may not have similar values in comparison with climatic observations and forecasts but regardless, one may expect to observe similar patterns to some extent.

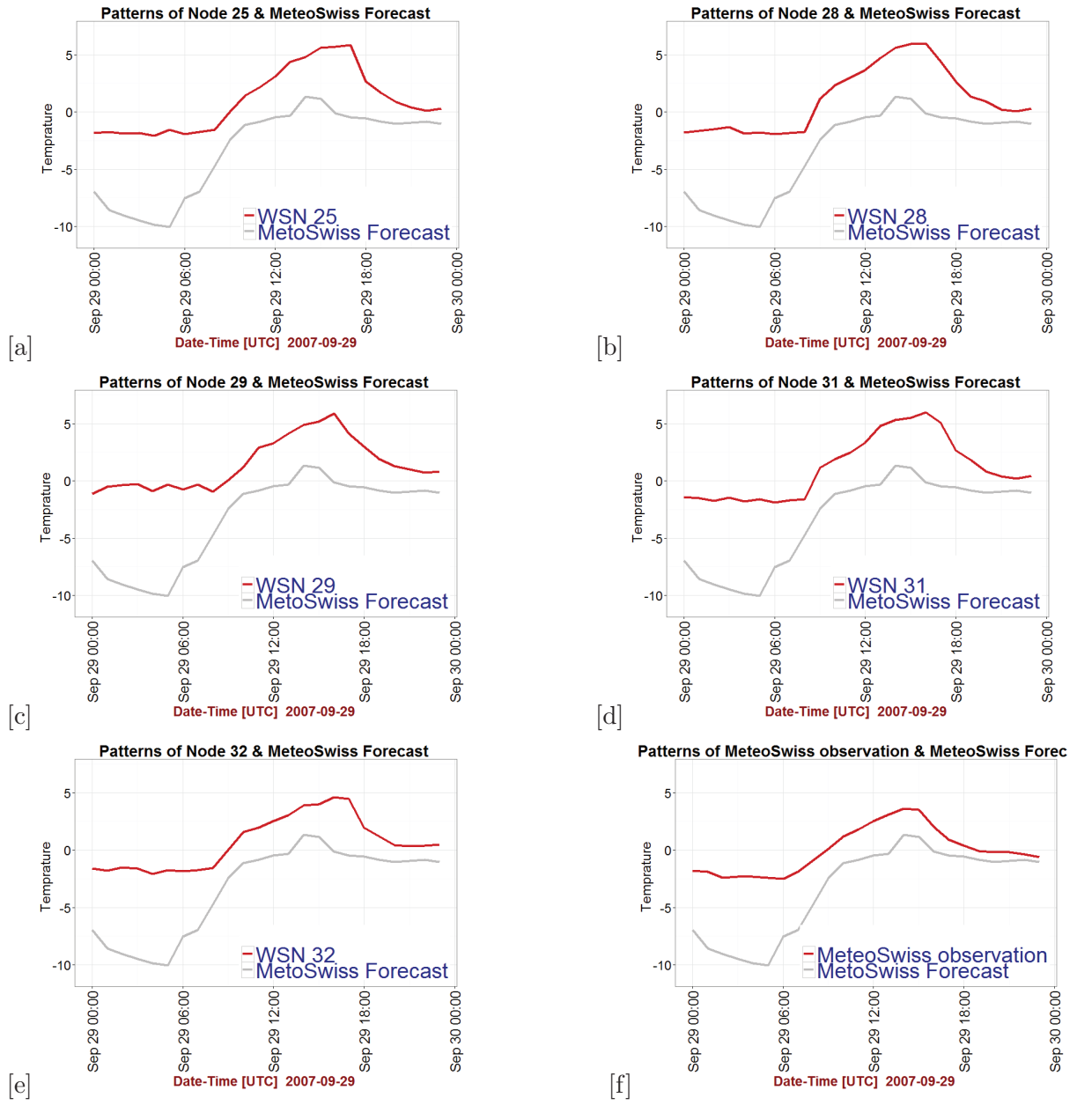


FIGURE 4.2: The patterns of wireless sensor network against the MeteoSwiss forecast for ambient temperature ($^{\circ}\text{C}$) on 2007-09-29

4.1.1 Temporal outlier detection using patterns

In order to identify temporal outliers, MeteoSwiss forecasts (MFs), provided by MeteoSwiss (2013), and WSN measurements are utilized to obtain temporal patterns and differences of

the patterns, present outliers. Additionally, the suitability of MeteoSwiss observations (MOs) for building the initial temporal model is evaluated. In other words, temporal correlation between consecutive values of MFs, MOs and WSN measurements are utilized to expose temporal patterns in order to identify temporal outliers.

Hourly forecasts up to 72 hours in advance, are provided by the numerical weather prediction model, COSMO-7 (Baldauf et al., 2011). COSMO models are described in Section 2.2.2. Moreover, the near surface temperatures are height-adjusted to the height of the study area (with a constant lapse rate), while the model level values are not. In addition due to local clouds [fog] and wind regimes in the complex topography cannot be resolved on a high spatial resolution by COSMO-7 model, while sensors of small cluster all within a grid mesh of 2×2 km. In order to overcome a problem of height difference between the meteorological station and WSNs stations, forecasts for four COSMO grid points around the study area were retrieved. Subsequently, the nearest grid point forecast to the WSN was utilized.

4.1.1.1 Pattern formation

As discussed, absolute values cannot address the outlier detection in this study and patterns are thus compared to identify outliers. Behavior of the temperature in a specific time interval is called a pattern. In this study, patterns are obtained by conjoining the values of temperature in consecutive time slices. In order to, form patterns WSN measurements and MeteoSwiss products should be in a similar time scales. While the WSN data were observed with two minutes frequency, MFs are hourly instantaneous values and MOs are hourly mean values.

MeteoSwiss forecasts are derived from instantaneous values at the full hour HH:00 and forecast model time step is 60s only and thus the real time resolution is about five minutes (MeteoSwiss, 2013). Subsequently, every hour of instantaneous WSN measurements are aggregated with observations of five minutes before and after and will be referred to as “instantaneous” values in the text.

MeteoSwiss observations are hourly mean values in 10-minute intervals for an hour HH = (HH-1):41 - HH:40. Hourly means were obtained by calculating the mean values of the following six time instants: (HH-1):50, HH:00, HH:10, HH:20, HH:30, HH:40.

WSNs measurements are aggregated by the mean function to obtain a consistent temporal resolution with respect to the MFs and MOs. Moreover, MFs and MOs were obtained by different approaches and separate WSN subsets are assigned to each of them. Thus, four types

of patterns exist for MFs, MOs and WSN (WSN measurements aggregated based on MF and MO).

4.1.1.2 Similarity assessment

Similarity assessment in this study was performed by calculating the slopes. For the similarity assessment slopes of WSNs measurements and MFs are compared as well as slopes of WSN and MOs. Slopes for WSNs measurements are calculated from differences between a current mean value and previous mean value within the same time period. On the other hand, slopes for MFs are computed by subtracting two successive hourly forecasts within similar time periods, since slopes of MOs are obtained by calculating the differences of two hourly mean sequences. Using the mean aggregation function to provide the similar temporal resolution in WSN yielded benefits such as reducing the effects of infrequent faults and reducing the effect of small fluctuations. Comparing patterns of WSN and MFs can also determine long term faults in WSNs. Moreover, using slopes to identify outliers ignores effects of vertical shifts, global scaling and shrinking (Suntinger et al., 2010) due to the complex topography and harsh condition of the study area.

4.1.1.3 Temporal outlier detection

Temporal outliers are those observations of WSN on a specific time slice where patterns of WSN and MeteoSwiss products show differences. Patterns are considered similar when the slopes of the MeteoSwiss products are placed within the tolerance of the WSN slope. In other words, if the slope of the MeteoSwiss products at specific hour lay outside the WSN slope tolerance, a subsequent relevant WSN observation of that hour is labeled as an outlier, otherwise labeled as normal. The tolerance is defined based on the accuracy of the wireless sensor device, presented in Table 3.1. Based on the accuracy of the device, the maximum and minimum range for each observation can be estimated. Thus, the tolerances are defined based on possible minimum and maximum variation for each WSN sequence.

4.2 Spatial outlier detection

In section 2.2.1 spatial modeling by using geostatistics is described. Geostatistical data analysis is used to model spatial correlation. This includes calculating the sample variogram to obtain spatial correlation, fitting a model to the sample variogram and using the model to make predictions at unsampled locations (Webster and Oliver, 2007).

4.2.1 Spatial outlier detection by using kriging

Spatial outliers are identified which uses the contextual information belonging to neighboring nodes. In order to identify spatial outliers, sample variograms should be calculated. Variogram modeling generally requires a sample size of at least 100 (Webster and Oliver, 2007). In this study WSN contains 23 nodes where six of nodes are faulty and cannot be used in variogram modeling. In other words, spatial variation of sensors data cannot be presented by 17 nodes since less than 50 locations are liable to sudden unpredictable change (Webster and Oliver, 2007). Hence, spatial variability analysis is performed using multiple measurements in time to estimate the variogram (Sterk and Stein, 1997). In other words, in order to overcome the aforementioned constraint the method of Sterk and Stein (1997) is adopted under the assumption of a similar spatial correlation over time. This implies that the spatial variability of WSN measurements during different time periods is characteristic of the same variogram model structure. Subsequently, the Equation 2.1 for the variogram is modified to Equation 4.1. A pair of observations with same separation vector h in different time intervals are grouped into similar classes.

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{j=1}^m \sum_{i=1}^{n_j(h)} (y(i, j) - y(i + h, j))^2 \quad (4.1)$$

where: m is the number of different time periods, h is the lag distance and $n_j(h)$ is the number of point pairs for each h at time period j . Spatial outliers are those observations that differed significantly from the predicted value.

Kriging applies a linear combination of observations to predict unbiased values in unsampled locations with minimum variance error. Observations and variogram model should be used to perform predictions. In order to identify spatial outliers, WSN observations and predictions should be compared and outliers are those observations that differ significantly. Spatial outliers are identified based on the confidence interval of the predictions by using the standard error of the predicted values and the coefficient for a given confidence level.

4.3 Event detection

In this research events are identified based on the correlation based neighbor-voting approach. When a sensor node identifies outliers it requires geographically correlated neighbors to help determine whether the outlier is a “fault” or an “event”. Outliers that represent a change in state of environment are spatially correlated for correlated nodes. Events are defined as those

observations where, temporal outliers occurred on at least half of the correlated sensor nodes, otherwise they are labeled as faults. Moreover, based on data exploration and pre-processing results, it is assumed that the probability of all sensor nodes being faulty simultaneously is very low. On the other hand, before voting for neighbor's outliers, the sensor node should be aware of the correlation rate of their adjacent nodes, as nodes only have interaction with adjacent correlated nodes.

4.4 Re-labeling

In Section 3.4 the methodologies underlying the labeling of the data are described. While that labeled data were derived from two-minute periods, the current methodology requires labeled data in hourly time resolution. In order to have hourly labeled data, re-labeling is performed based on the existing labeled dataset.

Re-labeling is accomplished with respect to the hourly instantaneous and hourly mean periods, described in Section 3.3. Both hourly instantaneous and hourly mean periods are re-labeled to extract the majority of the labels in the corresponding time periods, and are called “majority” approach in the text. Moreover, hourly instantaneous and hourly mean periods are used to re-label the data based on the time periods of hourly mean and hourly instantaneous and if any outliers exist in the related time period are classified within “minimum” approach, as it is referred in the text.

4.5 Evaluation in terms of accuracy and energy

In the context of WSNs, energy efficiency and accuracy are critical for performance evaluation given the technology requirements.

4.5.1 Accuracy of detected outliers

The accuracy of the outliers can be assessed with respect to the reference data. Assessments can be performed based on detection rate (DR) and false positive rate (FPR). DR presents the percentage of outliers that are detected correctly, and calculated using Equation 4.2, where OP is the number of correctly identified outliers and TP is a total number of detected outliers.

$$DR_{outlier} = (OP \div TP) \times 100 \quad (4.2)$$

On the other hand, FPR represents the percentage of non-outlier records that are detected as outliers and is defined by Equation 4.3, where OF is number of non-outlier records that identified as outliers and TN is total number of normal data.

$$FPR = (OF \div TN) \times 100 \quad (4.3)$$

4.5.2 Accuracy of detected events

The accuracy of the detected events can be evaluated with respect to the detection rate (DR_{event}) and false positive alarm (FPA). DR event presents the percentage of correctly detected events is described by Equation 4.4, where EP is the number of correctly identified events and TE is the total number of events.

$$DR_{event} = (EP \div TE) \times 100 \quad (4.4)$$

FPA presents the percentage of faults that are detected as events and is calculated by equation 4.5, where EF is number of faults that are incorrectly detected as events and TF is total number of faults.

$$FPA = (EF \div TF) \times 100 \quad (4.5)$$

4.5.3 Evaluating the spatial model

Cross-validation is known as the simplest and most widely used technique for estimating the errors of predictions (Webster and Oliver, 2007). Main metrics of cross-validation are, mean prediction error (MPE), which shows the quality of predictions in terms of bias and root mean square error (RMSE), for a model is a measure of accuracy.

4.5.4 Evaluation of complexity and energy efficiency

The complexity of applied method for event detection is assessed in terms of communication overhead, computation and memory complexity. Additionally, consumed energy for communication is investigated while, the main source of energy consumption in WSN is communication overhead.

4.6 Event characterizing

In order to retrieve detailed information about events and faults, the detected events' time scale is degraded. Events are investigated in finer resolution while detected in the hourly scale. Thus, in the presence of WSN measurements variance a tolerance is defined for labeling the events in degraded resolution. Consequently, based on two-minute sampling frequency of WSN data, nodes are re-labeled as "fault" or "event". Variance is calculated for the observations that are involved in the hour of occurred event. Subsequently, tolerance is defined based on the variance of the relevant observations and those WSN measurements that lie within the tolerance of "hourly value \pm variance", are labeled as event or otherwise as fault. In other words, by using the aforementioned procedure, corresponding observations of a specific time interval (i.e. a full instantaneous hour labeled as event) are re-labeled as event or fault.

Chapter 5

Results

5.1 Temporal outlier detection using patterns

This section presents result of temporal outlier detection in the presence of MeteoSwiss Forecasts (MFs), MeteoSwiss observations (MOs) and wireless sensor networks (WSNs) patterns. Temporal correlation between consecutive values of MFs, MOs and WSN measurements are utilized to establish temporal patterns in order to identify temporal outliers. Subsequently, similarity assessment of patterns is performed by determining differences of the patterns of wireless sensor nodes, with respect to the MFs and MOs. Because the time scales of the MOs and MFs are different, the WSN measurements should be aggregated to obtain a consistent time scale. The method of temporal outlier detection is described in section 4.1.1.

To make the WSN measurements consistent in term of time resolution WSN data on 2007–09–30 are aggregated with respect to the MeteoSwiss forecasts and MeteoSwiss observation specifications. MeteoSwiss forecasts are derived from instantaneous values at the full hour, while the forecast model time step is 60s only and thus the real time resolution is about five minutes. Consequently, every hour of instantaneous WSN measurements are aggregated with observations of five minutes before and after and are referred to as “WSN hourly instantaneous values” in the text and these values are presented by Table 5.1. Table 5.2 presents the results of the aggregated WSN measurements, with respect to the specification of MeteoSwiss observations that will be called “WSN hourly mean values” in the text. Additionally, hourly values for MeteoSwiss observations and forecasts for 2007–09–30 are shown in Table 5.3.

Patterns of hourly WSN values are compared against the MFs and MOs to identify temporal outliers. Figure 5.1 presents the patterns of instantaneous hourly values of WSN against the

TABLE 5.1: The hourly instantaneous values of WSN temperature observations on 2007–09–30 with respect to the specifications of MeteoSwiss forecasts ($^{\circ}\text{C}$)

Date Time	WSN25	WSN28	WSN29	WSN31	WSN32
29/09/2007 23:00	0.28	0.30	0.81	0.42	0.48
30/09/2007 00:00	0.21	0.31	0.53	0.36	0.26
30/09/2007 01:00	-0.26	-0.04	0.40	-0.20	-0.15
30/09/2007 02:00	-0.48	-0.43	0.04	-0.54	-0.32
30/09/2007 03:00	-0.71	-0.51	0.26	-0.70	-0.57
30/09/2007 04:00	-0.71	-0.65	0.27	-0.77	-0.53
30/09/2007 05:00	-0.50	-0.47	0.86	-0.59	-0.32
30/09/2007 06:00	-0.22	-0.20	0.58	-0.34	-0.07
30/09/2007 07:00	0.13	0.20	1.21	0.10	0.26
30/09/2007 08:00	0.57	0.49	1.34	0.28	0.59
30/09/2007 09:00	1.04	1.27	1.75	0.98	1.61
30/09/2007 10:00	1.09	1.56	1.58	1.34	1.27
30/09/2007 11:00	1.82	2.20	2.14	2.19	2.09
30/09/2007 12:00	4.57	4.34	4.15	4.78	2.96
30/09/2007 13:00	3.81	5.86	4.23	5.60	4.42
30/09/2007 14:00	7.40	5.34	4.89	7.46	6.67
30/09/2007 15:00	5.57	6.37	5.07	5.86	5.67
30/09/2007 16:00	5.40	5.68	5.98	4.83	5.20
30/09/2007 17:00	6.29	5.20	6.12	5.68	6.37
30/09/2007 18:00	4.97	4.58	5.23	4.48	4.78
30/09/2007 19:00	5.18	5.74	6.25	4.58	5.79
30/09/2007 20:00	2.38	2.63	3.47	2.75	2.77
30/09/2007 21:00	4.71	4.57	5.04	3.79	5.99
30/09/2007 22:00	5.37	5.47	5.61	4.70	5.29
30/09/2007 23:00	4.77	4.92	5.56	4.25	5.06

WSN25: WSN Node 25
 WSN28: WSN Node 28
 WSN29: WSN Node 29
 WSN31: WSN Node 31
 WSN32: WSN Node 32

MeteoSwiss forecasts on 2007–09–30. Figure 5.2 illustrates the patterns of hourly mean WSN values against the MeteoSwiss observations on 2007–09–30.

Temporal outliers are identified by analyzing the similarities in the slopes. The list of slopes that are calculated from WSN hourly instantaneous values are presented in Table 5.4 for 2007–09–30 and Table 5.5 presents slopes that are obtained from WSN hourly mean values on 2007–09–30.

Similarity assessment is applied on the slope values of WSN and MeteoSwiss products for corresponding time periods. Those observations of WSN are labeled as outliers in the case that the slope of the MeteoSwiss products do not place within the tolerance of the WSN slope. If the slope of MeteoSwiss products at specific hour lay outside the tolerance of the wireless sensor slopes, relevant WSN observations of that hour are considered outliers or otherwise labeled as normal. In this study ± 0.6 is selected as tolerance since the accuracy of the ambient temperature sensor is ± 0.3 .

A report of the detected outliers, with respect to the MeteoSwiss forecasts and WSN hourly instantaneous values on 2007–09–30 are revealed in Table 5.6. Table 5.7 presents the identified

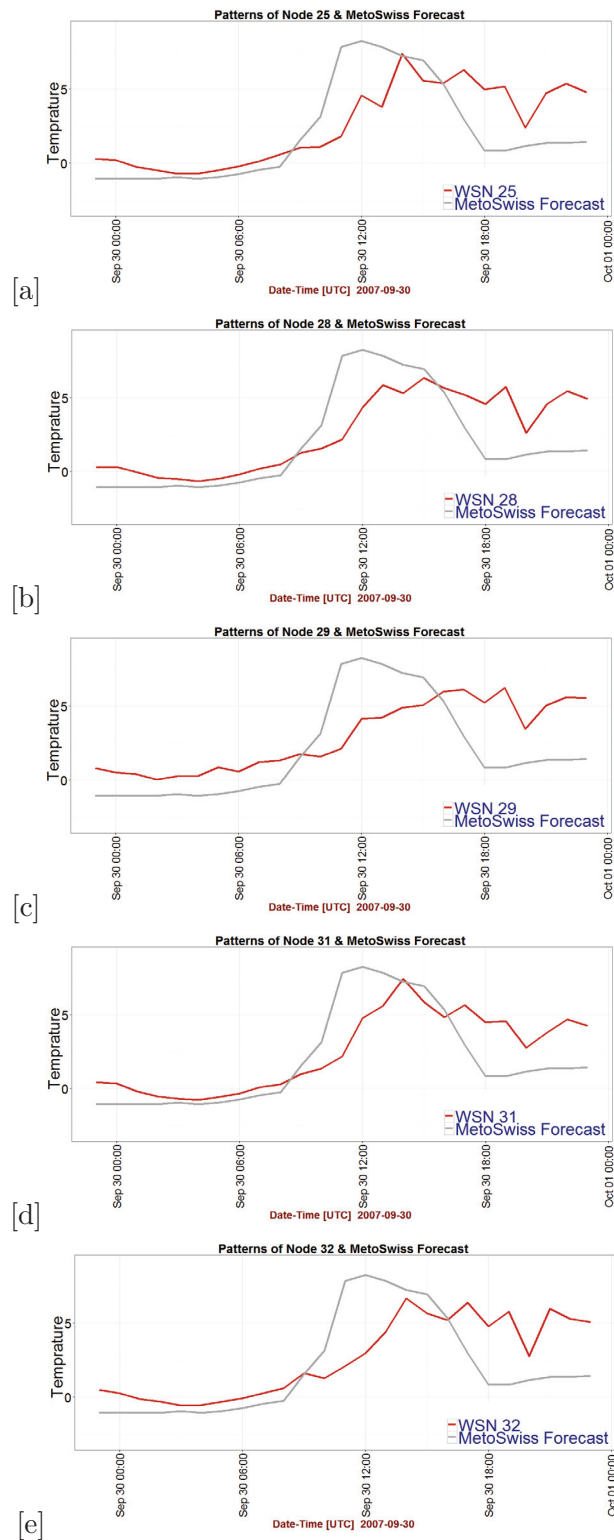


FIGURE 5.1: The patterns of instantaneous temperature values ($^{\circ}\text{C}$) of wireless sensor networks against the MeteoSwiss forecasts on 2007-09-30.

TABLE 5.2: The hourly mean values of WSN measurements on 2007–09–30 with respect to the specification of MeteoSwiss observations ($^{\circ}\text{C}$)

Date Time	WSN25	WSN28	WSN29	WSN31	WSN32
29/09/2007 23:00	0.26	0.35	0.69	0.38	0.38
30/09/2007 00:00	0.02	0.12	0.37	0.10	0.09
30/09/2007 01:00	-0.35	-0.21	0.14	-0.35	-0.20
30/09/2007 02:00	-0.42	-0.38	0.28	-0.49	-0.25
30/09/2007 03:00	-0.73	-0.62	0.25	-0.77	-0.57
30/09/2007 04:00	-0.65	-0.59	0.34	-0.72	-0.45
30/09/2007 05:00	-0.40	-0.34	0.59	-0.47	-0.21
30/09/2007 06:00	-0.09	-0.03	0.84	-0.17	0.10
30/09/2007 07:00	0.19	0.21	1.36	0.09	0.33
30/09/2007 08:00	0.64	0.74	1.45	0.49	0.78
30/09/2007 09:00	1.06	1.37	1.61	1.09	1.31
30/09/2007 10:00	1.27	1.85	1.91	1.44	1.35
30/09/2007 11:00	2.28	2.63	2.52	2.57	2.13
30/09/2007 12:00	4.17	4.19	4.01	4.21	3.44
30/09/2007 13:00	6.34	5.69	5.49	7.07	5.40
30/09/2007 14:00	6.57	5.48	5.20	6.14	5.60
30/09/2007 15:00	6.75	6.28	5.66	6.03	5.51
30/09/2007 16:00	5.83	5.60	6.07	5.45	5.92
30/09/2007 17:00	6.37	5.00	5.74	5.42	5.88
30/09/2007 18:00	4.60	4.54	5.30	4.58	4.97
30/09/2007 19:00	4.98	5.44	6.10	4.50	5.70
30/09/2007 20:00	2.66	2.49	4.30	2.44	2.69
30/09/2007 21:00	4.91	5.10	5.57	4.52	5.39
30/09/2007 22:00	5.05	5.44	5.88	4.58	5.22
30/09/2007 23:00	4.57	4.98	5.52	4.05	4.87

WSN25: WSN Node 25

WSN28: WSN Node 28

WSN29: WSN Node 29

WSN31: WSN Node 31

WSN32: WSN Node 32

temporal outliers, with respect to the MeteoSwiss observations and WSN hourly mean values on 2007–09–30. Likewise, Figure 5.3 shows the result of temporal outliers concerning the MF and WSN hourly instantaneous values on 2007–09–30. Figure 5.4 presents the result of temporal outliers with respect to the MeteoSwiss observations and WSN hourly mean values for small cluster of nodes on 2007–09–30.

TABLE 5.3: The hourly values of MeteoSwiss forecasts and MeteoSwiss observations on 2007-09-30 (°C)

Hour	MeteoSwiss Forecasts	MeteoSwiss Observations
29/09/2007 23:00	-1.05	-0.60
30/09/2007 00:00	-1.05	-0.60
30/09/2007 01:00	-1.05	-1.20
30/09/2007 02:00	-1.05	-1.00
30/09/2007 03:00	-0.95	-0.80
30/09/2007 04:00	-1.05	-0.40
30/09/2007 05:00	-0.95	-0.20
30/09/2007 06:00	-0.75	0.10
30/09/2007 07:00	-0.45	0.10
30/09/2007 08:00	-0.25	0.70
30/09/2007 09:00	1.55	1.00
30/09/2007 10:00	3.15	1.40
30/09/2007 11:00	7.85	2.30
30/09/2007 12:00	8.25	3.40
30/09/2007 13:00	7.85	3.50
30/09/2007 14:00	7.25	3.50
30/09/2007 15:00	6.95	4.20
30/09/2007 16:00	5.35	4.50
30/09/2007 17:00	2.95	4.80
30/09/2007 18:00	0.85	2.90
30/09/2007 19:00	0.85	4.20
30/09/2007 20:00	1.15	4.30
30/09/2007 21:00	1.35	3.80
30/09/2007 22:00	1.35	4.10
30/09/2007 23:00	1.45	4.00

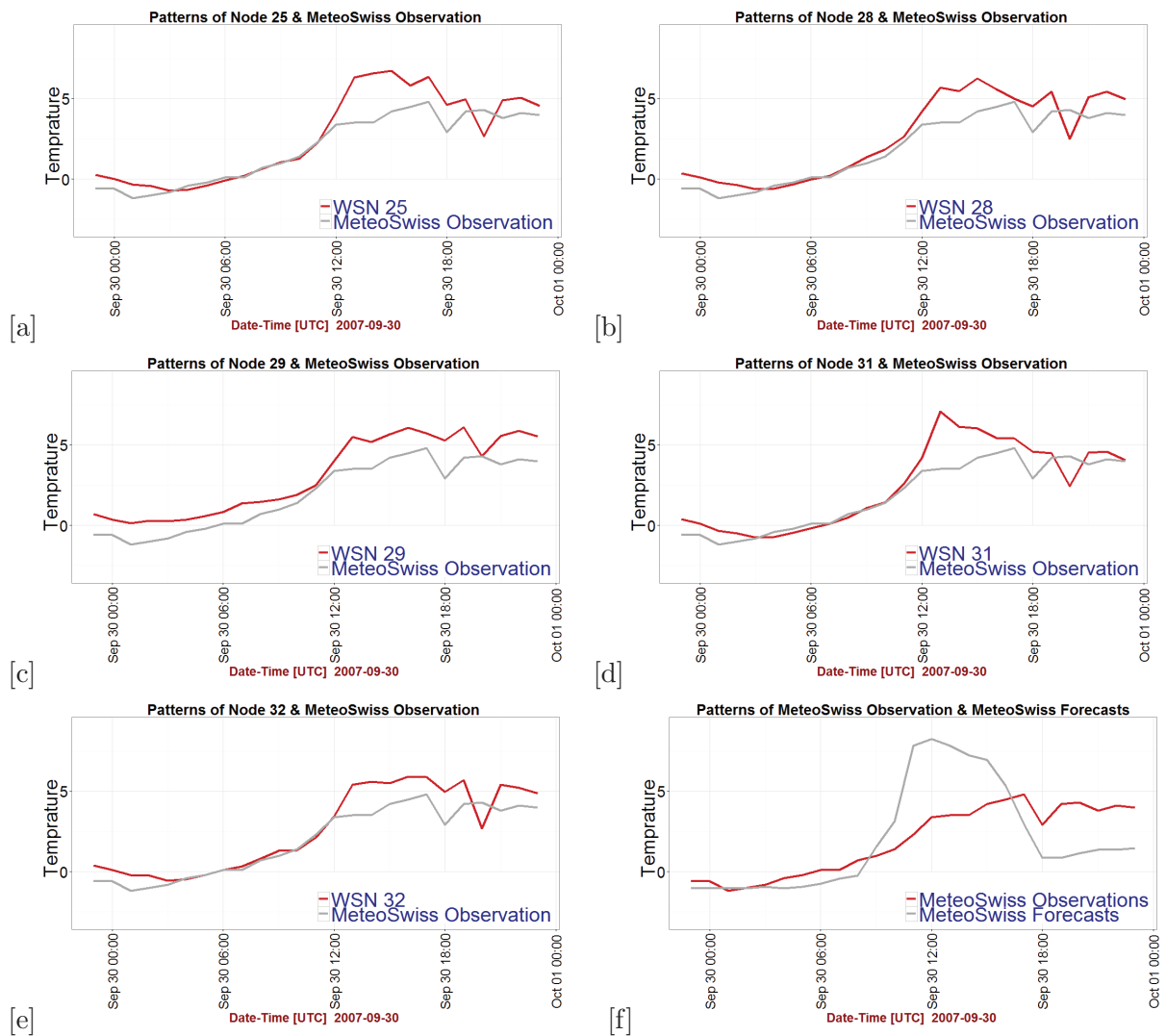


FIGURE 5.2: The patterns of hourly mean temperature values ($^{\circ}\text{C}$) of wireless sensor networks against the MeteoSwiss observation on 2007-09-30.

TABLE 5.4: The slopes calculated from hourly instantaneous values of WSN measurements for 2007-09-30

Date Time	WSN25	WSN28	WSN29	WSN31	WSN32
30/09/2007 0:00	-0.072	0.016	-0.278	-0.062	-0.214
30/09/2007 1:00	-0.472	-0.354	-0.132	-0.554	-0.412
30/09/2007 2:00	-0.216	-0.392	-0.358	-0.342	-0.174
30/09/2007 3:00	-0.228	-0.076	0.218	-0.158	-0.244
30/09/2007 4:00	-0.006	-0.136	0.008	-0.072	0.034
30/09/2007 5:00	0.216	0.172	0.594	0.180	0.208
30/09/2007 6:00	0.282	0.274	-0.276	0.246	0.252
30/09/2007 7:00	0.342	0.402	0.626	0.440	0.334
30/09/2007 8:00	0.448	0.288	0.128	0.184	0.330
30/09/2007 9:00	0.468	0.776	0.410	0.694	1.018
30/09/2007 10:00	0.044	0.294	-0.172	0.368	-0.342
30/09/2007 11:00	0.734	0.638	0.560	0.844	0.824
30/09/2007 12:00	2.750	2.140	2.016	2.590	0.864
30/09/2007 13:00	-0.762	1.518	0.074	0.822	1.462
30/09/2007 14:00	3.592	-0.520	0.666	1.858	2.250
30/09/2007 15:00	-1.828	1.036	0.174	-1.598	-0.998
30/09/2007 16:00	-0.170	-0.690	0.912	-1.028	-0.472
30/09/2007 17:00	0.886	-0.482	0.144	0.844	1.176
30/09/2007 18:00	-1.322	-0.620	-0.894	-1.192	-1.592
30/09/2007 19:00	0.214	1.162	1.026	0.096	1.012
30/09/2007 20:00	-2.800	-3.112	-2.788	-1.830	-3.020
30/09/2007 21:00	2.330	1.942	1.572	1.040	3.218
30/09/2007 22:00	0.656	0.898	0.574	0.906	-0.698
30/09/2007 23:00	-0.600	-0.550	-0.050	-0.450	-0.238

WSN25: WSN Node 25

WSN28: WSN Node 28

WSN29: WSN Node 29

WSN31: WSN Node 31

WSN32: WSN Node 32

TABLE 5.5: The slopes obtained from hourly mean values of WSN measurements, MeteoSwiss forecasts and MeteoSwiss observations on 2007–09–30

Date Time	WSN25	WSN28	WSN29	WSN31	WSN32	MO	MF
30/09/2007 00:00	-0.25	-0.23	-0.33	-0.28	-0.29	0.00	0.00
30/09/2007 01:00	-0.37	-0.33	-0.22	-0.45	-0.29	-0.60	0.00
30/09/2007 02:00	-0.07	-0.17	0.14	-0.13	-0.05	0.20	0.00
30/09/2007 03:00	-0.31	-0.25	-0.03	-0.28	-0.32	0.20	0.10
30/09/2007 04:00	0.08	0.03	0.09	0.05	0.11	0.40	-0.10
30/09/2007 05:00	0.25	0.25	0.25	0.25	0.24	0.20	0.10
30/09/2007 06:00	0.31	0.32	0.25	0.30	0.31	0.30	0.20
30/09/2007 07:00	0.28	0.24	0.52	0.26	0.23	0.00	0.30
30/09/2007 08:00	0.45	0.53	0.09	0.40	0.46	0.60	0.20
30/09/2007 09:00	0.42	0.63	0.16	0.59	0.53	0.30	1.80
30/09/2007 10:00	0.21	0.49	0.30	0.35	0.04	0.40	1.60
30/09/2007 11:00	1.01	0.78	0.61	1.13	0.78	0.90	4.70
30/09/2007 12:00	1.89	1.56	1.49	1.64	1.31	1.10	0.40
30/09/2007 13:00	2.17	1.50	1.48	2.86	1.96	0.10	-0.40
30/09/2007 14:00	0.23	-0.21	-0.30	-0.93	0.20	0.00	-0.60
30/09/2007 15:00	0.18	0.80	0.47	-0.11	-0.09	0.70	-0.30
30/09/2007 16:00	-0.92	-0.67	0.41	-0.59	0.41	0.30	-1.60
30/09/2007 17:00	0.54	-0.60	-0.34	-0.02	-0.04	0.30	-2.40
30/09/2007 18:00	-1.76	-0.47	-0.44	-0.84	-0.90	-1.90	-2.10
30/09/2007 19:00	0.38	0.90	0.81	-0.08	0.73	1.30	0.00
30/09/2007 20:00	-2.32	-2.95	-1.81	-2.06	-3.01	0.10	0.30
30/09/2007 21:00	2.25	2.61	1.27	2.08	2.70	-0.50	0.20
30/09/2007 22:00	0.14	0.34	0.31	0.06	-0.18	0.30	0.00
30/09/2007 23:00	-0.48	-0.46	-0.36	-0.53	-0.35	-0.10	0.10

WSN25: WSN Node 25

WSN28: WSN Node 28

WSN29: WSN Node 29

WSN31: WSN Node 31

WSN32: WSN Node 32

MO: MeteoSwiss observation

MF: MeteoSwiss forecast

TABLE 5.6: The temporal outliers presentation for hourly instantaneous values that are detected with respect to the MeteoSwiss forecasts on 2007–09–30 where, ✓ symbol presents normal measurements

Date — Time	Node ID				
	25	28	29	31	32
30/09/2007 00:00	✓	✓	✓	✓	✓
30/09/2007 01:00	✓	✓	✓	✓	✓
30/09/2007 02:00	✓	✓	✓	✓	✓
30/09/2007 03:00	✓	✓	✓	✓	✓
30/09/2007 04:00	✓	✓	✓	✓	✓
30/09/2007 05:00	✓	✓	✓	✓	✓
30/09/2007 06:00	✓	✓	✓	✓	✓
30/09/2007 07:00	✓	✓	✓	✓	✓
30/09/2007 08:00	✓	✓	✓	✓	✓
30/09/2007 09:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 10:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 11:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 12:00	Outlier	Outlier	Outlier	Outlier	✓
30/09/2007 13:00	✓	Outlier	✓	Outlier	Outlier
30/09/2007 14:00	Outlier	✓	Outlier	Outlier	Outlier
30/09/2007 15:00	Outlier	Outlier	✓	Outlier	Outlier
30/09/2007 16:00	Outlier	Outlier	Outlier	✓	Outlier
30/09/2007 17:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 18:00	Outlier	Outlier	Outlier	Outlier	✓
30/09/2007 19:00	✓	Outlier	Outlier	✓	Outlier
30/09/2007 20:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 21:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 22:00	Outlier	Outlier	✓	Outlier	Outlier
30/09/2007 23:00	Outlier	Outlier	✓	✓	✓

TABLE 5.7: The temporal outliers of hourly means values that are detected with respect to the MeteoSwiss observation on 2007–09–30, where ✓ presents normal measurements

Date — Time	Node ID				
	25	28	29	31	32
30/09/2007 00:00	✓	✓	✓	✓	✓
30/09/2007 01:00	✓	✓	✓	✓	✓
30/09/2007 02:00	✓	✓	✓	✓	✓
30/09/2007 03:00	✓	✓	✓	✓	✓
30/09/2007 04:00	✓	✓	✓	✓	✓
30/09/2007 05:00	✓	✓	✓	✓	✓
30/09/2007 06:00	✓	✓	✓	✓	✓
30/09/2007 07:00	✓	✓	✓	✓	✓
30/09/2007 08:00	✓	✓	✓	✓	✓
30/09/2007 09:00	✓	✓	✓	✓	✓
30/09/2007 10:00	✓	✓	✓	✓	✓
30/09/2007 11:00	✓	✓	✓	✓	✓
30/09/2007 12:00	Outlier	✓	✓	✓	✓
30/09/2007 13:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 14:00	✓	✓	✓	Outlier	✓
30/09/2007 15:00	✓	✓	✓	Outlier	Outlier
30/09/2007 16:00	Outlier	Outlier	✓	Outlier	✓
30/09/2007 17:00	✓	Outlier	Outlier	✓	✓
30/09/2007 18:00	✓	Outlier	Outlier	Outlier	Outlier
30/09/2007 19:00	Outlier	✓	✓	Outlier	✓
30/09/2007 20:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 21:00	Outlier	Outlier	Outlier	Outlier	Outlier
30/09/2007 22:00	✓	✓	✓	✓	✓
30/09/2007 23:00	✓	✓	✓	✓	✓

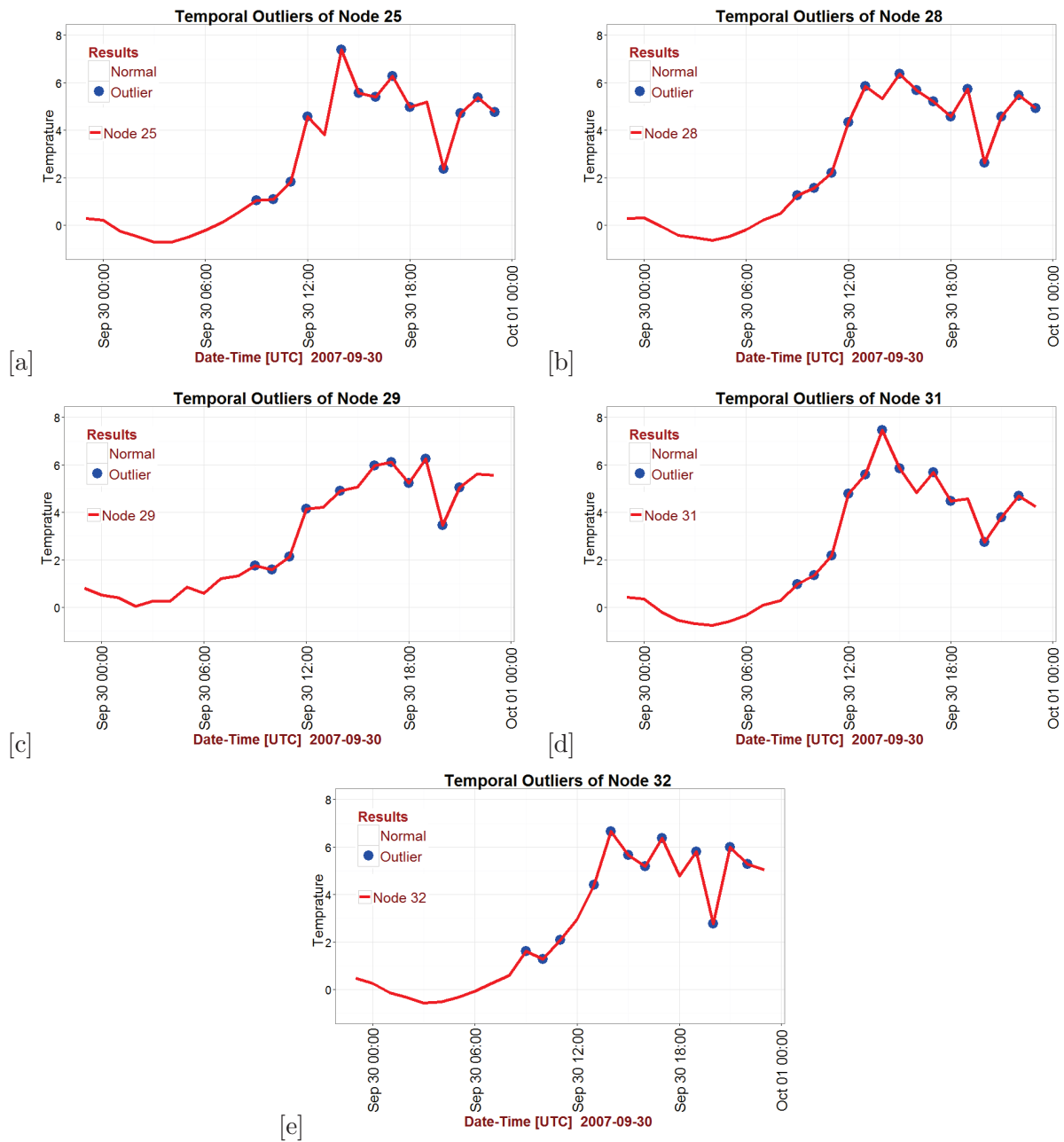


FIGURE 5.3: The temporal outlier detection results for WSN hourly instantaneous values with respect to the MeteoSwiss forecast on 2007-09-30

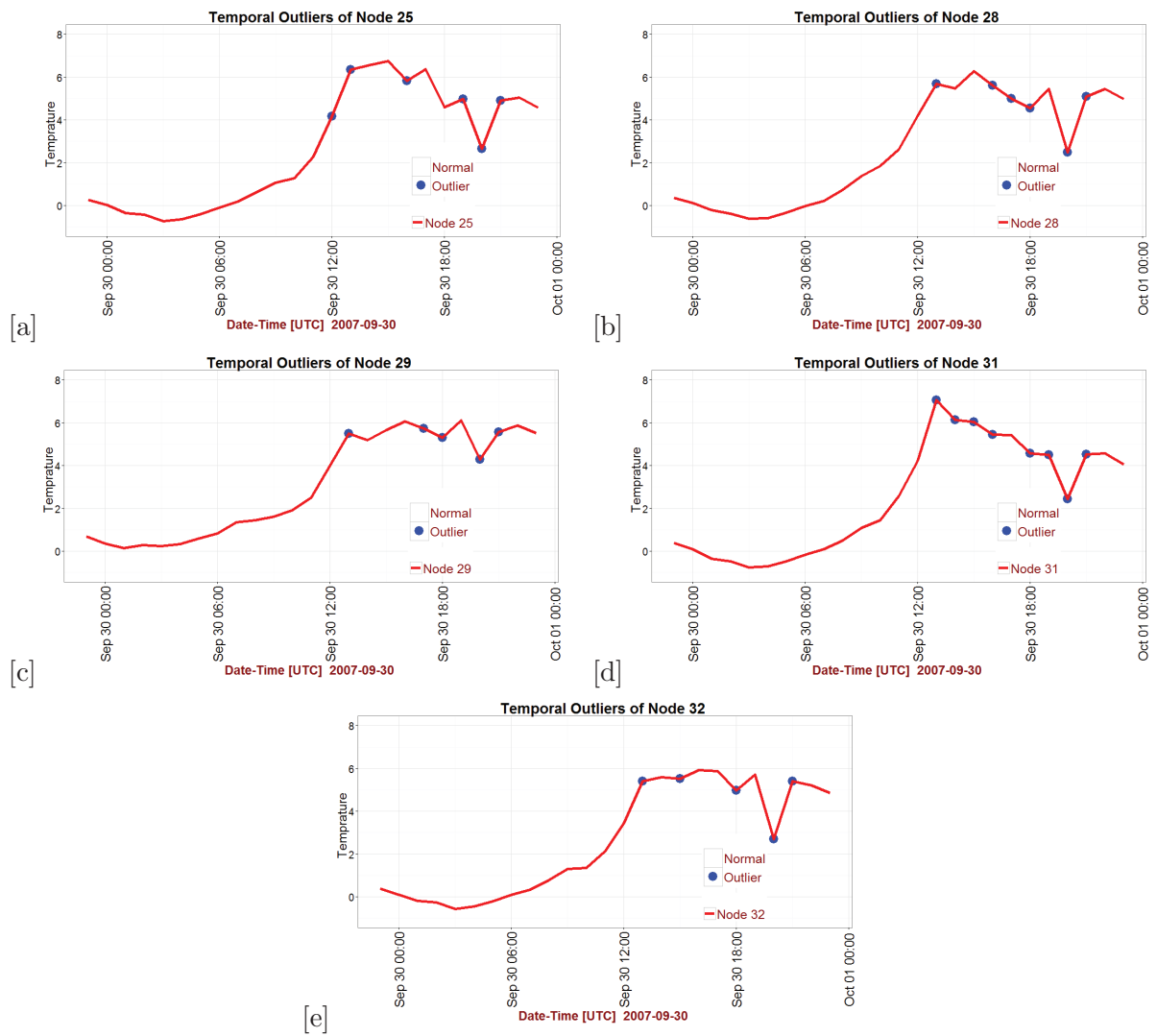


FIGURE 5.4: The temporal outlier detection results for WSN hourly means values with respect to the MeteoSwiss observations on 2007-09-30

5.2 Spatial outlier detection using kriging

Spatial outlier detection method uses observations of spatial neighbors for prediction and then for outlier detection. In this study, the variogram is obtained from Zhang et al. (2012) for the same dataset. The calibrated sample variogram and the fitted exponential model are shown in Figure 5.5(a) for 2007–09–30. The properties of the fitted model from Zhang et al. (2012) are presented in Table 5.8. Moreover, the same variogram is calibrated for predicting the hourly instantaneous values for the 2007–09–29 data and Figure 5.5(b) illustrates the corresponding calibrated variogram. Subsequently, predictions can be obtained with respect to the fitted model where Table 5.9 presents the results of predictions for the small cluster on 2007–09–30 and Table 5.10 shows the prediction results for 2007–09–29.

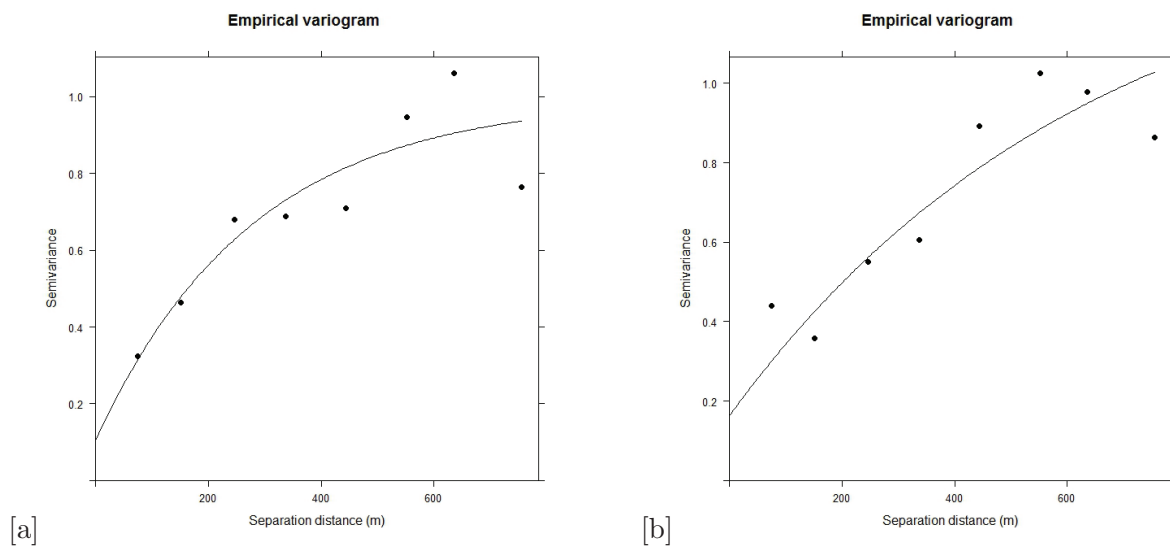


FIGURE 5.5: Illustration of variogram and the exponential fitted model where (a) variogram for 2007–09–30 and (b) variogram for 2007–09–29

TABLE 5.8: The properties of the fitted exponential model

Partial Sill	Nugget	Range
1	0.2	550

Spatial outliers are those observations that lay outside the 95% confidence interval of the predicted values in the spatial domain. Results show that there are not any spatial outliers for the nodes of small cluster on 2007–09–30 and node 25 is identified as a spatial outlier on 07:00, 2007–09–29.

TABLE 5.9: The instantaneous values of WSN temperature measurements ($^{\circ}\text{C}$) and correspondent predictions on 2007-09-30, where “O” presents the instantaneous values of temperature and “P” presents the predicted values ($^{\circ}\text{C}$)

Time — Date	Node 25		Node 28		Node 29		Node 31		Node 32	
	O	P	O	P	O	P	O	P	O	P
30/09/2007 00:00	0.42	0.21	0.49	0.31	0.43	0.53	0.30	0.36	0.44	0.26
30/09/2007 01:00	-0.05	-0.26	0.16	-0.04	0.01	0.40	-0.09	-0.20	-0.05	-0.15
30/09/2007 02:00	-0.30	-0.48	-0.14	-0.43	-0.28	0.04	-0.36	-0.54	-0.26	-0.32
30/09/2007 03:00	-0.45	-0.71	-0.12	-0.51	-0.40	0.26	-0.49	-0.70	-0.43	-0.57
30/09/2007 04:00	-0.45	-0.71	-0.12	-0.65	-0.46	0.27	-0.52	-0.77	-0.42	-0.53
30/09/2007 05:00	-0.24	-0.50	0.26	-0.47	-0.30	0.86	-0.28	-0.59	-0.26	-0.32
30/09/2007 06:00	-0.08	-0.22	0.21	-0.20	-0.10	0.58	-0.08	-0.34	-0.09	-0.07
30/09/2007 07:00	0.32	0.13	0.72	0.20	0.28	1.21	0.31	0.10	0.27	0.26
30/09/2007 08:00	0.55	0.57	0.90	0.49	0.55	1.34	0.65	0.28	0.58	0.59
30/09/2007 09:00	1.25	1.04	1.42	1.27	1.17	1.75	1.28	0.98	1.05	1.61
30/09/2007 10:00	1.35	1.09	1.45	1.56	1.46	1.58	1.32	1.34	1.27	1.27
30/09/2007 11:00	2.16	1.82	2.15	2.20	2.17	2.14	2.02	2.19	2.05	2.09
30/09/2007 12:00	3.91	4.57	4.13	4.34	4.10	4.15	4.13	4.78	4.10	2.96
30/09/2007 13:00	4.86	3.81	4.58	5.86	5.16	4.23	4.54	5.60	4.34	4.42
30/09/2007 14:00	6.58	7.40	5.84	5.34	5.63	4.89	6.31	7.46	6.43	6.67
30/09/2007 15:00	5.61	5.57	5.34	6.37	5.93	5.07	5.74	5.86	5.52	5.67
30/09/2007 16:00	5.03	5.40	5.42	5.68	5.34	5.98	5.49	4.83	5.08	5.20
30/09/2007 17:00	5.92	6.29	5.97	5.20	5.46	6.12	5.95	5.68	5.87	6.37
30/09/2007 18:00	4.75	4.97	4.97	4.58	4.75	5.23	4.86	4.48	4.91	4.78
30/09/2007 19:00	5.27	5.18	5.63	5.74	5.56	6.25	5.59	4.58	5.26	5.79
30/09/2007 20:00	2.86	2.38	3.13	2.63	2.70	3.47	2.66	2.75	2.67	2.77
30/09/2007 21:00	4.69	4.71	4.69	4.57	4.52	5.04	4.93	3.79	4.48	5.99
30/09/2007 22:00	5.01	5.37	5.24	5.47	5.28	5.61	5.40	4.7	5.13	5.29
30/09/2007 23:00	4.72	4.77	5.05	4.92	4.84	5.56	4.96	4.25	4.73	5.06

TABLE 5.10: The instantaneous values of WSN temperature measurements and correspondent predictions on 2007-09-29, where “O” presents the instantaneous values of temperature and “P” presents the predicted values ($^{\circ}\text{C}$)

Time — Date	Node 25		Node 28		Node 29		Node 31		Node 32	
	O	P	O	P	O	P	O	P	O	P
30/09/2007 00:00	-1.86	-1.47	-1.79	-1.32	-1.13	-1.68	-1.42	-1.67	-1.61	-1.67
30/09/2007 01:00	-1.73	-1.50	-1.67	-1.07	-0.53	-1.62	-1.53	-1.54	-1.79	-1.63
30/09/2007 02:00	-1.88	-1.46	-1.49	-1.02	-0.37	-1.54	-1.75	-1.46	-1.51	-1.68
30/09/2007 03:00	-1.83	-1.36	-1.32	-0.91	-0.27	-1.42	-1.47	-1.41	-1.60	-1.59
30/09/2007 04:00	-2.09	-1.76	-1.87	-1.37	-0.89	-1.82	-1.79	-1.84	-2.08	-1.88
30/09/2007 05:00	-1.54	-1.50	-1.77	-0.97	-0.32	-1.66	-1.62	-1.46	-1.76	-1.58
30/09/2007 06:00	-1.95	-1.69	-1.94	-1.28	-0.74	-1.83	-1.89	-1.73	-1.85	-1.81
30/09/2007 07:00	-1.73	-1.55	-1.86	-1.02	-0.33	-1.76	-1.70	-1.56	-1.75	-1.70
30/09/2007 08:00	-1.54	-1.55	-1.74	-1.27	-0.92	-1.68	-1.63	-1.51	-1.57	-1.65
30/09/2007 09:00	-0.00	0.51	1.14	0.32	0.05	0.79	1.14	0.30	0.00	0.32
30/09/2007 10:00	1.44	1.63	2.30	1.41	1.18	1.92	1.89	1.64	1.55	1.54
30/09/2007 11:00	2.18	2.37	3.00	2.58	2.89	2.69	2.44	2.47	1.96	2.41
30/09/2007 12:00	3.10	3.03	3.64	3.15	3.24	3.41	3.29	3.15	2.51	3.19
30/09/2007 13:00	4.34	4.04	4.71	4.19	4.12	4.54	4.77	4.14	3.04	4.33
30/09/2007 14:00	4.77	4.74	5.58	4.87	4.88	5.29	5.30	4.83	3.86	4.95
30/09/2007 15:00	5.61	4.95	5.92	5.16	5.17	5.68	5.50	5.30	3.98	5.47
30/09/2007 16:00	5.70	5.46	5.96	5.72	5.85	5.86	5.96	5.57	4.58	5.73
30/09/2007 17:00	5.82	4.73	4.41	4.60	4.13	4.78	5.08	4.91	4.47	5.26
30/09/2007 18:00	2.65	2.51	2.63	2.78	2.96	2.73	2.66	2.57	1.93	2.75
30/09/2007 19:00	1.68	1.58	1.34	1.77	1.88	1.51	1.82	1.52	1.20	1.64
30/09/2007 20:00	0.86	0.78	0.88	1.03	1.29	0.90	0.81	0.85	0.44	0.88
30/09/2007 21:00	0.38	0.43	0.19	0.69	1.01	0.31	0.37	0.41	0.31	0.37
30/09/2007 22:00	0.11	0.32	0.07	0.48	0.71	0.16	0.16	0.25	0.38	0.17
30/09/2007 23:00	0.28	0.49	0.29	0.62	0.80	0.37	0.42	0.40	0.47	0.36

5.2.1 Evaluation of spatial model

Cross-validation is applied in order to estimate the prediction errors. The mean prediction error (MPE) and root mean square error (RMSE) values are calculated to investigate the bias and accuracy of the predictions. For the spatial model, MPE is -0.033 and RMSE is 0.759 on 2007–09–30. Moreover, the cross-validation evaluation of the predictions presents a more reliable spatial model on 2007–09–29, because the model is calculated based on the 2007–09–29 data where the MPE for relevant day is 0.013 and the RMSE is 0.597. The low MPE revealed that the model is unbiased and the low RMSE indicates that the model is accurate.

Aforementioned results reveal that assumption of approximately constant spatial structure over time is correct, while a same variogram results show accurate predictions on 2007–09–29 and 2007–09–30. By assuming the approximately same spatial correlation structure over time, temperature in this study should be irrespective of time scale, wind, precipitation and humidity. From the results in Section 5.2.1 one can see that by using the variogram averaging method the same variogram can work properly for different time scales and is further proof of approximately constant spatial structure over time. In addition, by applying the method of variogram averaging and using pre-calculated variograms the computation rate can be reduced considerably. The primary, drawback of the spatial correlation based outlier detection is that it cannot identify temporal outliers that are happened in multiple nodes, simultaneously.

5.3 Event detection

The method of event detection is explained in Section 4.3. Table 5.11 shows the amount and direction of correlation between WSN small cluster nodes. Consequently, all the nodes can participate in event detection.

TABLE 5.11: The correlation between wireless sensor nodes [25–28–29–31–32]

	Node Name				
	25	28	29	31	32
Node 25	1.00	0.96	0.92	0.97	0.97
Node 28	0.96	1.00	0.93	0.97	0.96
Node 29	0.92	0.93	1.00	0.92	0.94
Node 31	0.97	0.97	0.92	1.00	0.96
Node 32	0.97	0.96	0.94	0.96	1.00

Events on 2007–09–30 are identified with respect to the result of temporal outliers for MeteoSwiss forecasts (Table 5.6) and MeteoSwiss observations (Table 5.7). Subsequently, correlation based neighbor-voting is applied to identify events that are presented in Table 5.12 for MeteoSwiss forecasts and Table 5.13 MeteoSwiss observations.

TABLE 5.12: The result of detected events for MeteoSwiss forecasts on 2007–09–30 where, ✓ symbol presents normal measurements and ☹ symbol presents events

Date — Time	Node ID				
	25	28	29	31	32
30/09/2007 00:00	✓	✓	✓	✓	✓
30/09/2007 01:00	✓	✓	✓	✓	✓
30/09/2007 02:00	✓	✓	✓	✓	✓
30/09/2007 03:00	✓	✓	✓	✓	✓
30/09/2007 04:00	✓	✓	✓	✓	✓
30/09/2007 05:00	✓	✓	✓	✓	✓
30/09/2007 06:00	✓	✓	✓	✓	✓
30/09/2007 07:00	✓	✓	✓	✓	✓
30/09/2007 08:00	✓	✓	✓	✓	✓
30/09/2007 09:00	☹	☹	☹	☹	☹
30/09/2007 10:00	☹	☹	☹	☹	☹
30/09/2007 11:00	☹	☹	☹	☹	☹
30/09/2007 12:00	☹	☹	☹	☹	✓
30/09/2007 13:00	✓	☹	✓	☹	☹
30/09/2007 14:00	☹	✓	☹	☹	☹
30/09/2007 15:00	☹	☹	✓	☹	☹
30/09/2007 16:00	☹	☹	☹	✓	☹
30/09/2007 17:00	☹	☹	☹	☹	☹
30/09/2007 18:00	☹	☹	☹	☹	✓
30/09/2007 19:00	✓	☹	☹	✓	☹
30/09/2007 20:00	☹	☹	☹	☹	☹
30/09/2007 21:00	☹	☹	☹	☹	☹
30/09/2007 22:00	☹	☹	✓	☹	☹
30/09/2007 23:00	Fault	Fault	✓	✓	✓

TABLE 5.13: The result of detected events for MeteoSwiss observations on 2007–09–30 where, ✓ symbol presents normal measurements and ☁ symbol presents events

Date — Time	Node ID				
	25	28	29	31	32
30/09/2007 00:00	✓	✓	✓	✓	✓
30/09/2007 01:00	✓	✓	✓	✓	✓
30/09/2007 02:00	✓	✓	✓	✓	✓
30/09/2007 03:00	✓	✓	✓	✓	✓
30/09/2007 04:00	✓	✓	✓	✓	✓
30/09/2007 05:00	✓	✓	✓	✓	✓
30/09/2007 06:00	✓	✓	✓	✓	✓
30/09/2007 07:00	✓	✓	✓	✓	✓
30/09/2007 08:00	✓	✓	✓	✓	✓
30/09/2007 09:00	✓	✓	✓	✓	✓
30/09/2007 10:00	✓	✓	✓	✓	✓
30/09/2007 11:00	✓	✓	✓	✓	✓
30/09/2007 12:00	Fault	✓	✓	✓	✓
30/09/2007 13:00	☁	☁	☁	☁	☁
30/09/2007 14:00	✓	✓	✓	Fault	✓
30/09/2007 15:00	✓	✓	✓	Fault	Fault
30/09/2007 16:00	☁	☁	✓	☁	✓
30/09/2007 17:00	✓	Fault	Fault	✓	✓
30/09/2007 18:00	✓	☁	☁	☁	☁
30/09/2007 19:00	Fault	✓	✓	Fault	✓
30/09/2007 20:00	☁	☁	☁	☁	☁
30/09/2007 21:00	☁	☁	☁	☁	☁
30/09/2007 22:00	✓	✓	✓	✓	✓
30/09/2007 23:00	✓	✓	✓	✓	✓

5.4 Results of re-labeling

In Section 4.4 re-labeling approaches are described based on running average-based, Mahalanobis distance-based and density based labeling techniques.

Re-labeling in temporal domain is performed for three labeling technique for small cluster of nodes on 06:00-14:00, 2007-09-30. Table 5.14 presents the result of re-labeling for temporal density based labeling technique. Result of re-labeling for temporal Mahalanobis distance-based labeling are illustrated in Table 5.15. Table 5.16 shows the result of temporal re-labeling for running average-based labeling technique.

TABLE 5.14: The result of re-labeling for temporal density based labeling technique on 2007-09-30

Node ID	A	B	C	D
25	-	13:00	-	-
28	-	-	-	-
29	-	-	-	-
31	-	13:00	-	-
32	-	-	-	-

A: Hourly mean Majority
 B: Hourly mean Minimum
 C: Instantaneous Majority
 D: Instantaneous Minimum

TABLE 5.15: The result of re-labeling for temporal Mahalanobis distance-based labeling technique on 2007-09-30

Node ID	A	B	C	D
25	-	13:00	-	-
28	-	-	-	-
29	-	13:00	-	-
31	-	13:00	-	14:00
32	-	13:00	-	14:00

A: Hourly mean Majority
 B: Hourly mean Minimum
 C: Instantaneous Majority
 D: Instantaneous Minimum

TABLE 5.16: The result of re-labeling for temporal running average-based labeling technique on 2007-09-30

Node ID	A	B	C	D
25	-	12:00, 13:00, 14:00	-	14:00
28	-	12:00, 13:00, 14:00	-	12:00, 14:00
29	-	12:00, 13:00	-	-
31	-	12:00, 13:00, 14:00	-	14:00
32	-	13:00, 14:00	-	14:00

A: Hourly mean Majority
 B: Hourly mean Minimum
 C: Instantaneous Majority
 D: Instantaneous Minimum

5.5 Temporal outlier detection accuracy

This section presents the results of temporal outlier detection in terms of DR and FPR by using re-labeled data as is described in Section 4.4.

Table 5.17 shows the results of DR and FPR for temporal outliers using the three labeling techniques for WSN hourly instantaneous values with respect to the MeteoSwiss forecasts on 2007–09–30, by using the minimum approach of re-labeling for the hourly instantaneous period.

Table 5.18 shows the DR and FPR by using the minimum approach of re-labeling for the hourly mean period. Moreover, based on relabeled data there are not any outliers for hourly mean or instantaneous periods in the majority approach, thus the DRs are calculated as zero percent and FPRs are 100%.

TABLE 5.17: The temporal outlier detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach of instantaneous period

Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPR%	DR%	FPR%	DR%	FPR%
25	0	100	0	100	20	100
28	0	100	0	100	20	100
29	0	100	0	100	0	100
31	0	100	16	100	16	100
32	0	100	20	100	20	100

TABLE 5.18: The temporal outlier detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach hourly mean period

Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPR%	DR%	FPR%	DR%	FPR%
25	0	100	0	100	40	60
28	0	100	0	100	40	75
29	0	100	0	100	20	100
31	16	100	16	100	50	100
32	0	100	20	100	40	75

Table 5.19 shows the DR and FPR for temporal outliers using the three labeling techniques for WSN hourly mean values, based on MeteoSwiss observations and with respect to the majority re-labeled data on the hourly the instantaneous period. Table 5.20 presents the DR and FPR for temporal outliers of WSN hourly mean values based on the minimum approach of re-labeled data with respect to the hourly time period. Moreover, based on re-labeled data there are not any outliers for hourly mean or instantaneous majority approach, again the DRs are calculated as zero percent and FPRs are 100%.

TABLE 5.19: The temporal outlier detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 based on the minimum approach of re-labeled data and hourly instantaneous time period

Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPR%	DR%	FPR%	DR%	FPR%
25	0	28	0	28	0	28
28	0	12	0	12	0	12
29	0	100	0	100	0	100
31	0	28	50	14	50	14
32	0	14	0	14	0	14

TABLE 5.20: The temporal outlier detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 based on the minimum approach of re-labeled data and hourly mean time period

Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPR%	DR%	FPR%	DR%	FPR%
25	50	14	50	14	100	0
28	0	12	0	12	0	12
29	0	12	100	0	100	0
31	50	14	50	14	100	0
32	0	12	100	0	100	0

5.6 Event detection accuracy

The accuracy of the detected events is assessed in terms of DR and FPA based on the equations in Section 4.5.2. Events are assigned in the re-labeled data based on correlation based neighbor-voting, where the majority of nodes represent outliers in same time instant. Events are evaluated based on hourly mean and minimum approach of re-labeled data, as there were no events within re-labeled data based on the majority approach and instantaneous time intervals; these are presented in Section 5.4. Table 5.21 presents the accuracy of hourly mean events with respect to the MeteoSwiss observations and Table 5.22 shows the accuracy of hourly instantaneous events with respect to the MFs.

TABLE 5.21: The events detection accuracy of hourly mean values with respect to the MOs on 2007–09–30 by using minimum approach of hourly period

Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPA%	DR%	FPA%	DR%	FPA%
25	0	0	100	0	100	100
28	0	0	100	0	100	0
29	0	0	100	0	100	0
31	0	0	100	0	100	0
32	0	0	100	0	100	0



TABLE 5.22: The event detection accuracy of hourly instantaneous values with respect to the MFs on 2007–09–30 by using minimum approach hourly instantaneous period
























































































































Node ID	Density-based		Mahalanobis distance-based		Running average-based	
	DR%	FPA%	DR%	FPA%	DR%	FPA%
25	0	0	0	0	50	0
28	0	0	20	0	40	0
29	0	0	0	0	40	100
31	0	0	16	0	60	100
32	0	0	20	0	20	0

5.7 Event characterizing

Spatial and temporal developments of events were presented in Table 5.12 for MeteoSwiss forecasts and in Table 5.13 for MeteoSwiss observation in hourly scales. Events of MeteoSwiss observations are chosen to further explore information, as it presents a higher detection rate. Events that were detected based on MeteoSwiss observations are utilized to perform degradation of time resolution, where outliers for all the nodes were detected on 13:00 and degradation is applied to re-label the data in finer resolution on 12:50~13:40, based on MeteoSwiss specifications for observations.

The event characterizing is explored in Section 4.6. Table 5.23 delineates events of small cluster that occurred in all nodes on 12:50~13:40. Node 25 presented a non-stop event for whole the hour, while the other nodes' events disappeared once at least. The event started from nodes 25, 28, 29, 32 and after six minutes was observed in node 31. The event lasted for one hour and finished at the same time in all the nodes.

TABLE 5.23: Events characteristics in fine resolution based on MeteoSwiss observations on 12:50~13:40, 2007-09-30 where,  presents events and  symbol presents faults

Date — Time	Node ID				
	25	28	29	31	32
30/09/2007 12:50					
30/09/2007 12:52					
30/09/2007 12:54					
30/09/2007 12:56					
30/09/2007 12:58					
30/09/2007 13:00					
30/09/2007 13:02					
30/09/2007 13:04					
30/09/2007 13:06					
30/09/2007 13:08					
30/09/2007 13:10					
30/09/2007 13:12					
30/09/2007 13:14					
30/09/2007 13:16					
30/09/2007 13:18					
30/09/2007 13:20					
30/09/2007 13:24					
30/09/2007 13:26					
30/09/2007 13:28					
30/09/2007 13:30					
30/09/2007 13:32					
30/09/2007 13:34					
30/09/2007 13:36					
30/09/2007 13:38					
30/09/2007 13:40					

5.8 Model complexity and energy efficiency

Energy efficient WSNs should be assessed in terms of energy complexity as well as accuracy. In order to evaluate complexity of the method, the detection technique should be assessed in terms of communication overhead, computation and memory complexity.

To detect events, temporal outlier detection should be performed, firstly. While the temporal detection is locally accomplished in each node, no communication overhead is needed. The computational complexity in temporal outlier detection should be considered in two conditions: using MeteoSwiss forecasts and using the temporal model that is built based on MeteoSwiss observations. The computation complexity for MeteoSwiss forecasts can be connived due this fact no much computation is needed. For MeteoSwiss observations and using them to build a initial temporal model, computation complexity depends on fitting and updating the model. The memory complexity differed for MeteoSwiss observations and forecasts while it is mainly about the storing of observations of WSN and MeteoSwiss forecast and it is negligible. For event detection method, each node sends the message to the neighbors when an outlier is observed. The maximum communication overhead for each node is $O(n.v.m)$, where n is the total number of correlated neighbors, v is the number of variables and m is the number of detected temporal outliers.

The main source of energy consumption in WSNs is communication. Thus, this research only focuses on communication overhead to evaluate the energy efficiency. To calculate the energy consumption in terms of communication, properties of utilized device (Instruments, 2007) are presented in Table 5.24.

TABLE 5.24: The properties of single-chip 2.4 GHz IEEE 802.15.4 compliant RF transceiver applied in wireless sensor networks (Instruments, 2007)

Mac layer	IEEE 802.15.4
speed: Transmit bit rate	250 kbps
Operation frequency	2.4 GHz
Packet size	128 bytes
Radio model	TI CC2420
Transmit current at 0dBm	17.4 mA
Transmit current at -25dBm	8.5 mA
Receive current	18.8 mA
Supply voltage	(1.6-2.0 V)

The energy consumption in communication overhead is calculated by (Torres, 2006):

$$E = P \times T, \quad P = I \times V, \quad T = \frac{\text{Packet Size}}{\text{Transmit bit rate}} \quad (5.1)$$

Subsequently, required power for sending each message is $34.8 \times 10^{-3} \text{W}$ since maximum current is considered for the network due to the distances of the nodes. Required power for receiving each message is calculated as $37.6 \times 10^{-3} \text{W}$. By considering the possible maximum communication overhead, it is assumed within each hour there would be outlier in all the nodes and total power for sending and receiving four messages is obtained as $289.6 \times 10^{-3} \text{W}$. Besides, T is calculated as $4 \times 10^{-3} \text{S}$ with respect to the packet size and transmit bit rate. Finally, consumed energy for the assumed condition is resulted as $1158.4 \times 10^{-6} \text{J}$. Additionally, initial energy for each node considered as five J which is a common setup.

Chapter 6

Discussion

6.1 Data pre-processing

Cleaning the data from gross errors and observations that may have resulted from faulty or blocked nodes is vital as they may corrupt the detection methods, especially those detection methods that use the data to train. However, identifying the faults and errors from different sources is only part of the problem and dealing with these errors and faults is critical. Due to inexpensive and imprecise nature of sensors, faults happen regularly and they usually occur in successive sequences. For example, dead band errors usually happen in a time period. Generally, based on the degree of the fault and the source of the fault one may use or ignore them. Using fault observations introduces an error to the detection method, especially in the case of modeling. However, ignoring the faults may cause a potential loss of information. Faults can happen in either short or long time periods and they may also occur in single or multiple nodes. It should draw attention that the choice of applied methods for replacing the faults are highly affected by several elements such as: type of sensed variable, geographic properties of study site, spatial distribution of neighboring sensors and time of the observations (Schlatter, 1975; Bennett et al., 1984; Thiebaut and Pedder, 1987). This current research focuses on identifying the faults, however dealing with them is beyond the scope. Faults that have occurred in multiple adjacent nodes and within a long time period cannot be replaced easily by methods such as, spatial interpolation weighting methods, for the estimation of ignored values to create the serially complete data, or temporal correlation methods that use neighboring observations in time to estimate the ignored values. Moreover, missing values should be added to aforementioned sources of errors and faults, which impose more complexity for replacing the data.

6.2 Re-Labeling

While the initial labeled dataset was recorded in two-minute frequency, data are re-labeled to obtain hourly labels, with respect to the instantaneous and hourly mean time periods. Therefore, outliers that are not located in interested time frames based on MeteoSwiss specifications, were ignored in labeling results.

The main concern about these labeling techniques is that they are inherently just other outlier detection methods and they include some degree of uncertainty. The running average-based technique identifies outliers based on adjacent observations in time and does not look at the entire dataset. To adequately assess a method, a non-similar methods' results should be utilized. Thus, the running average-based is not a reliable method for evaluating the result of applied temporal outlier detection since both utilize a similar approach for outlier detection. On the other hand, Mahalanobis distance-based and density-based techniques label outliers in the presence of the entire dataset and do not consider the temporal order (Zhang et al., 2012). Thus, Mahalanobis distance-based and density-based techniques can present more reliable results for temporal outlier detection. Additionally, the running average-based technique works with respect to the surrounding values in time, providing further evidence that it is also not a proper method for spatial outlier detection. Finally, the choice for a particular labeling technique may lead to different result for outlier and event detection method thus the focus should be on the applicability as well as the accuracy of the methodology.

6.3 Temporal outlier detection

Detecting outliers based on temporal correlation of WSNs measurements can result high detection rates (Zhang et al., 2012). Due to insufficient historical data, temporal correlation models cannot usually be calculated accurately in the absence of diurnal and seasonal patterns. Insufficient numbers of historical data can also be resulted because of removing the faults due to inexpensive sensor devices, communication problems, duplicate values, gross errors, dead band errors and missed values. Moreover, there are some applications that need temporary deployments, so large amount of historical data would not be expected. Hence, a lack of historical data is usually a primary concern in the WSN area.

In order to overcome the aforementioned constrains, this research introduces quite novel approach for temporal outlier detection. Nowadays, climatic stations and forecasts are widely available. Thus, it may be assumed, in the presence of climatic stations and forecasts, the

temporal correlation based outlier detection can be performed. Climatic stations that utilize accurate sensors and sense multiple variables can provide two general types of products, raw observations and forecasts. This research proposed, using climatic observations as historical data to build the temporal correlation model or using climatic forecasts in WSN nodes, to identify temporal outliers. Suitability of these products is assessed in terms of detection rate and false positive rate.

The following issues should be considered when choosing the aforementioned products to apply in the WSN area. The raw observations and forecasts that are provided by climatic stations cannot be as local as WSNs for the entire network, which is a key point of WSNs existence. On the other hand, WSN devices are not as accurate as climatic sensors, since climatic sensors are not faced with energy and expense constrains. Consequently, even when operating in a same site the sensors of WSN and climatic stations may observe different values so that comparing absolute values may not result in accurate outliers. However, similar patterns between WSN nodes and climatic station or climatic forecast grid points may exist. To temporal outlier detection climatic forecasts and observations should present approximately similar patterns with respect to the WSN, where not only the distance from WSN but also the similarities of the environments are extremely crucial.

In this research, the slopes approach was utilized to assess the similarity of patterns, where energy hunger methods cannot be applied. The slope approach, in this case, is fitting as it can ignore the effect of vertical shifts of values and global scaling. Additionally, using slope can reduce the effect of localization for climatic products.

Results provided by Table 5.17, Table 5.18, Table 5.19 and Table 5.20 reveal that accurate detections resulted from climatic observations (MeteoSwiss observations) in the presence of high DR and low FPR. It should be noted that the MeteoSwiss sensing station is closer to the WSN in comparison with MeteoSwiss forecast grid point (Figure 3.5). Consequently, MeteoSwiss patterns are more similar to WSN patterns in terms of shape and values, as are illustrated in Figure 5.1 and Figure 5.2.

MeteoSwiss forecasts resulted from 7×7 km grid mesh that were produced by the COSMO-7 model. The selected MeteoSwiss forecast grid point is the most representative grid point of the MeteoSwiss sensing station because the orography is more similar than at other grid points and it is the nearest grid point to the WSN. While using COSMO-2 models would produce more localized forecasts the COSMO-2 was not operational on 2007 and thus the COSMO-7 forecast was utilized. As it is illustrated in Figure 5.2, MeteoSwiss forecasts, MeteoSwiss

observations and WSN follow quite similar patterns in shape and value on 00:00~08:00 and the main differences appear on 10:00~15:00. To answer which elements cause the differences, relative humidity of MeteoSwiss forecasts and observations are explored by Table 6.1. As the table reveals, from 10:00~15:00 forecasts and observations for humidity differed approximately 10%~15% as did the temperatures in the same period. It can be concluded that forecasts were made based on an event from 10:00~15:00, that did not happen, yet caused an increase of temperature. Thus, false positives that were resulted on 10:00~15:00, based on MeteoSwiss forecasts, presents an uncertainty of forecasts regarding the assuring occurrence of events (Table 5.6). Comparing Table 5.6 and Table 5.7 shows that partially similar outliers were observed from 16:00~21:00. It should draw attention that a high false positive rate for MeteoSwiss forecasts maybe caused by of an advanced time period of forecasts and, forecasts within shorter time spans should include less uncertainty. In this study forecasts were provided 36 hours in advanced from 00:00~24:00, 2007–09–30, conceptually, forecasts for 2007–09–29 should be more precise and an evidence is provided by Figure 4.2 where quite more similar patterns are represented for WSN, MFs and MOs in shape. Moreover, by using forecast values in WSNs outliers and events can be detected in an unsupervised fashion and semantics can be assigned to the detected events based on occurring or not occurring an event. Additionally, forecast variable as well as other climatic contextual information can be introduced to the WSN.

MeteoSwiss observations presented more reliable DR and FPR with respect to the forecasts. Subsequently, MeteoSwiss observations can potentially be utilized to build the temporal correlation models. While resulted predictions do not present a completely localized representation of WSN measurements, their degree of localization can be improved by updating and calibrating the model via WSN measurements.

6.4 Spatial outlier detection

To overcome small size of samples, variogram averaging method of Sterk and Stein (1997) was utilized under the assumption of similar spatial correlation over time. Based on the results presented in Section 5.2.1, the assumption was justified while, the spatial predictions for two different days present accurate and similar results. Besides, utilizing a same variogram for different days could reduce the consumption of computation energy in nodes.

Spatial outlier detection by using spatial correlation could not identify outliers, efficiently. The main drawback of the applied method is, it cannot detect outliers when they happen in multiple

TABLE 6.1: Hourly values of air humidity 2 meter above ground for MeteoSwiss forecasts and MeteoSwiss observations on 2007–09–30.

Date Time	A	B
30/09/2007 00:00	89.2	97.9
30/09/2007 01:00	95.8	100
30/09/2007 02:00	100	99.5
30/09/2007 03:00	100	100
30/09/2007 04:00	97.9	99.3
30/09/2007 05:00	98.6	99.5
30/09/2007 06:00	97.5	98.8
30/09/2007 07:00	98.5	99.3
30/09/2007 08:00	99.1	97.1
30/09/2007 09:00	96.5	97.6
30/09/2007 10:00	93.1	100
30/09/2007 11:00	80.8	98.7
30/09/2007 12:00	80.3	95.7
30/09/2007 13:00	81.3	97.1
30/09/2007 14:00	83.3	98.9
30/09/2007 15:00	84.7	95.8
30/09/2007 16:00	91.4	90.6
30/09/2007 17:00	99.1	84.8
30/09/2007 18:00	100	97.4
30/09/2007 19:00	100	85.1
30/09/2007 20:00	100	79.8
30/09/2007 21:00	100	84.8
30/09/2007 22:00	100	86.0
30/09/2007 23:00	100	85.2

A: Air humidity 2 m above ground, MeteoSwiss forecasts (%)

B: Air humidity 2 m above ground, MeteoSwiss observations (%)

correlated nodes since events are usually geographically correlated. Thus, the method can identify faults while, faults are not spatially correlated.

6.5 Event detection

The applied event detection method, can distinguish long-term errors from events by using contextual information provided by MeteoSwiss forecasts since those faults do not happen simultaneously in multiple nodes. Events can be distinguished from long-term faults by the proposed method while, those event detection methods that utilize temporal correlation models cannot distinguish events from long-term faults. Furthermore, due to imprecise nature of wireless sensors that are applied in WSNs context, it is likely to occur faults in multiple nodes simultaneously. Faults in multiple nodes can be distinguished by performing the event detection method for multiple variables of a network.

In this study due to this fact MeteoSwiss products were in hourly scale thus detected events were also in hourly resolution. Based on the users requirements and interested events hourly detection and subsequently proposed method can be (near) real-time.

Generally, there is no prior knowledge about the upcoming events thus detection methods should be unsupervised. On the other hand, retrieving information about the spatial and temporal properties of detected events and type of events are essential. Thus, events are investigated in finer resolution to retrieve detailed information about temporal and spatial properties of events.

6.6 Accuracy of detected outliers and events

Low detection rate of MeteoSwiss forecasts does not warn the method is inapplicable. In the absence of reference data inappropriate dataset was applied for method evaluation while the reference data should be a ground truth or chosen from other methods results with respect to the current dataset and methodology. It should also be noted that, detected outliers and events are matter of definition and based on the application domain and end user requirements the method can detect events and outliers, specifically. In other words, each definition of outlier and event can result different events and distinct interpretations can be made for outliers and events. Additionally, low detection rate for MeteoSwiss forecasts also can be resulted because of different definitions for outliers and events of labeled data and proposed method.

6.7 Model complexity and energy efficiency

Applying energy efficient methods for outlier and event detection increases the networks lifetime which is a prerequisite for an accurate detection. Lack of energy in nodes can lead to inaccurate observations or even can cause incomplete data transmission. The main concern over energy efficiency and complexity are expressed for communication overhead. The proposed temporal outlier detection is energy efficient in terms of communication overhead. The communication complexity depends on the transmission rate since applied method for temporal outlier detection identifies outliers locally. Typically, being energy efficient is described by comparing different methods and one of them is considered as energy efficient method, eventually. In this research, to detect events correlated adjacent nodes behavior is studied and uncorrelated nodes behavior is ignored which can lead to considerable reduction of transmission. Definitely, the less energy consumption results more energy efficient method. Furthermore, a trade-off exists between energy consumption and accuracy. The conventional event detection methods that utilize neighbors information do not consider the correlation approach which leads to using only effective information. Thus, proposed method is energy efficient while transmission was limited to only effective and correlated nodes.

Chapter 7

Conclusion

The main objective of this research is to develop energy efficient and accurate techniques to identify outliers and events with respect to the spatial and temporal properties of the wireless sensor networks (WSNs) data. In order to address the aforementioned objective, a number of specific objectives and corresponding research questions were formulated in Section 1.2.2. Specific objectives and the research questions are addressed through the literature review, achieved results and discussion section of the thesis.

7.1 To define outliers and events

Temporal outliers were defined as those WSN observations that differed from MeteoSwiss forecasts or observations, in terms of patterns. Spatial outliers were defined as those observations that were different from the predicted values provided by spatial correlation modeling. Subsequently, events were defined by the correlation based neighbor-voting approach and temporal outliers were labeled as events where the majority of the correlated sensors simultaneously represented the outlier. Thus, co-occurring of temporal outliers in the same time instant, within the majority of the correlated nodes, represented change of outliers to events.

7.2 To detect an event

Temporal and spatial correlation that exist between WSN measurements were applied to perform event detection.

Based on temporal correlation properties, observations close to each other in time are more similar. In this research, temporal model was not introduced from WSN. Temporal correlation between consecutive values of MFs and WSN measurements are utilized to establish temporal patterns in order to identify temporal outliers. Similarity between temporal patterns of WSN and MeteoSwiss products were assessed to identify temporal outliers. While a sufficient amount of historical data was not available, MeteoSwiss forecast values were applied to identify temporal outliers. Moreover, MeteoSwiss observations were utilized to evaluate the capability of them to perform the initial temporal models for WSN.

Based on regionalized theory, observations close to each other in space are more similar. Geostatistical analysis was applied to model the spatial correlation and to identify spatial outliers. Variogram averaging method (Sterk and Stein, 1997) which extends the spatial variability analysis to space–time domain, was utilized in the absence of minimum required sample size. Spatial outliers were defined as those observations that deviated from predicted values with respect to the tolerance obtained from the confidence interval. Spatial correlation based models did not show high detection rate in dense and limited areas where high correlation between nodes exists. Thus, temporal outliers that occurred in multiple nodes in the same time instant cannot be identified as outliers using spatial correlation based outlier detection methods. In this research, only information concerning nodes correlation was applied to identify events. Consequently, events were defined based on the co-existence of temporal outliers in correlated spatial neighbor nodes within the same time frame.

7.3 To characterize an event

Detected events can be characterized in terms of spatial extent and temporal evolution. Table 5.23 shows the events of small cluster occurred in all nodes on 12:50-13:40. Node 25 represented a non-stop event for whole the hour, whereas the other nodes events disappeared once at least. The event started from nodes 25, 28, 29, 32 and then after six minutes observed in node 31. The event lasted one hour and finished at the same time in all the nodes.

7.4 To evaluate the detected outliers and events

In the absence of hourly labeled data, re-labeling was performed on three labeling technique: density based, Mahalanobis distance-based and running average-based. Existing labels of Zhang

(2010) were re-labeled based on two time periods (hourly mean and instantaneous values) and two aggregation approaches (majority and minimum) that were explained in Section 4.4.

Accuracy of temporal outliers was assessed with respect to the re-labeled data and the different criteria and results are presented in Section 5.5. Temporal outliers that resulted from MeteoSwiss forecasts did not present high detection rates (DR) and low false positive rates (FPR). Maximum accuracy for temporal outlier detection of MeteoSwiss forecasts was obtained by using re-labeling data from the minimum approach during the hourly mean period which was 38% DR and 82% FPR. On the other hand, accurate detection was shown by the minimum approach of re-labeled data during the hourly mean time period with 80% DR and 2% FPR, using the running average labeling method for temporal outlier detection of MeteoSwiss observations.

Accuracy of events was assessed with respect to the re-labeled data and assigning event labels based on the correlation based neighbor-voting approach. Subsequently, maximum accuracy for events was determined by MeteoSwiss forecasts which observed a 42% DR and 40% false positive alarm (FPA) for the running average-based technique. Additionally, maximum accuracy for events determined by MeteoSwiss observation was obtained by 100% DR and 0% FPA for Mahalanobis distance-based labeling technique.

Complexity and energy efficiency of applied method were assessed. In WSNs the majority of energy is consumed by communication. Within the confines of this research, communication is needed in the presence of outliers and for correlated nodes. Based on the analysis performed in Section 5.8, this research can conclude that the applied method is energy friendly.

7.5 Recommendation

Developing a labeling technique that can distinguish between different types of outliers and identify events is proposed for future work since conventional methods are mainly another outlier detection methods. Choosing an adequate labeling technique for each method and dataset is essential and requires further work. In this research, temporal outlier detection is performed based on comparing the patterns in similar time instantaneous. In large networks events may occur in different time instantaneous, thus temporal outlier detection should also be sensitive for searching the patterns in different time periods. Performing the event detection based on building the temporal correlation model that is built by climatic observations cannot be implemented due to the time constraint in this research and is an open topic for future research.

Bibliography

- Aggarwal, C. and Yu, P. (2001). Outlier detection for high dimensional data. *ACM Sigmod Record*, 30(2):37–46.
- Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422.
- Albanese, A., Pal, S. K., and Petrosino, A. (2012). Rough sets, kernel set and spatio-temporal outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints):1.
- Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the cosmo model: description and sensitivities. *Monthly Weather Review*, 139(12):3887–3905.
- Barnett, V. and Lewis, T. (1984). Outliers in statistical data. *Applied Probability and Statistics, Wiley Series in Probability and Mathematical Statistics*. Chichester: Wiley, 1984, 2nd ed., 1.
- Bennett, R., Haining, R., and Griffith, D. (1984). The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, 74(1):138–156.
- Bettencourt, L., Hagberg, A., and Larkey, L. (2007). Separating the wheat from the chaff: Practical anomaly detection schemes in ecological applications of distributed sensor networks. *Distributed Computing in Sensor Systems*, pages 223–239.
- Brownrigg, R. (2012). maps: Draw geographical maps. <http://cran.r-project.org/web/packages/maps/index.html/>.
- Conway, J., Eddelbuettel, D., Nishiyama, T., Kumar Prayaga, S., and Tiffin, N. (2007). RPostgreSQL. <http://cran.r-project.org/web/packages/RPostgreSQL/index.html>. [Online; accessed 2013/01/30].

- Dereszynski, E. W. and Dietterich, T. G. (2011). Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Trans. Sen. Netw.*, 8(1):3:1–3:36.
- Díaz-Ramírez, A., Tafoya, L., Atempa, J., and Mejía-Alvarez, P. (2012). Wireless sensor networks and fusion information methods for forest fire detection. *Procedia Technology*, 3:69–79.
- Ding, M., Chen, D., Xing, K., and Cheng, X. (2005). Localized fault-tolerant event boundary detection in sensor networks. In *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM*, volume 2, pages 902–913, Miami, FL, USA. IEEE.
- Estrin, D., Govindan, R., Heidemann, J., and Kumar, S. (1999). Next century challenges: Scalable coordination in sensor networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, MobiCom '99*, pages 263–270, New York, NY, USA. ACM.
- Ingelrest, F., Barrenetxea, G., Schaefer, G., Vetterli, M., Couach, O., and Parlange, M. (2010). Sensorscope: Application-specific sensor network for environmental monitoring. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):17.
- Instruments, T. (2007). CC2420 Single-Chip 2.4 GHz IEEE 802.15. 4 Compliant and ZigBee Ready RF Transceiver.
- Jin, G. and Nittel, S. (2006). Ned: An efficient noise-tolerant event and event boundary detection algorithm in wireless sensor networks. In *7th International Conference on Mobile Data Management MDM 2006.*, pages 153–153, Nara, Japan. IEEE.
- Kahle, D. and Wickham, H. (2012). ggmap: A package for spatial visualization with google maps and openstreetmap. <http://cran.r-project.org/web/packages/ggmap/index.html/>.
- Kapitanova, K. and Son, S. (2010). Event detection in wireless sensor networks. *Work-in-Progress Proceedings*, page 37.
- Kaufmann, P. (2008). Association of surface stations to NWP model grid points. Technical report, MeteoSwiss, Zurich, Switzerland.
- Krishnamachari, B. and Iyengar, S. (2004). Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. *IEEE Transactions on Computers*, 53(3):241–250.

- Lewis, F. L. (2004). Wireless sensor networks. *Smart Environments: Technologies, Protocols, and Applications*, pages 11–46.
- Liu, J., Chu, M., Reich, J., and Zhao, F. (2003). State-centric programming for sensor-actuator network systems. *Pervasive Computing, IEEE*, 2(4):50–62.
- Ma, S., Wang, J., Liu, Z., and Jiang, H. (2013). Density-based distributed elliptical anomaly detection in wireless sensor networks. *Applied Mechanics and Materials*, 249:226–230.
- McIlroy, D. (2003). mapproj: Map projections. <http://cran.r-project.org/web/packages/mapproj/index.html>.
- MeteoSwiss (2013). Federal department of home affairs (FDHA), federal office of meteorology and climatology meteoswiss. www.meteoswiss.ch. [Online; accessed 2013/01/10].
- Ogbagabriel, T. (2012). Statistically-based event detection using wireless sensor networks. Master’s thesis, University of Twente.
- Onoz, B. and OGUZ, B. (2003). Assessment of outliers in statistical data analysis. *Integrated Technologies for Environmental Monitoring and Information Production*, pages 173–180.
- Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers Geosciences*, 30:683–691.
- Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. <http://CRAN.R-project.org/doc/Rnews/>.
- PostgreSQL (2013). <http://www.postgresql.org>. [Online; accessed 2013/01/20].
- Ribeiro Jr, P. J. and Diggle, P. J. (2001). geoR: a package for geostatistical analysis. <http://CRAN.R-project.org/doc/Rnews/>. ISSN 1609-3631.
- Schlatter, T. (1975). Some experiments with a multivariate statistical objective analysis scheme. *Mon. Wea. Rev.*, 103(3):246–257.
- Sensorscope (2007). Sensor networks for environmental monitoring. <http://lcav.epfl.ch/page-86035-en.html>. Grand-St-Bernard Deployment.
- Shahid, N. and Naqvi, I. (2011). Energy efficient outlier detection in wsns based on temporal and attribute correlations. In *7th International Conference on Emerging Technologies (ICET)*, pages 1–6, Islamabad, Pakistan. IEEE.

- Shahid, N., Naqvi, I., and Qaisar, S. (2012). Quarter-sphere svm: Attribute and spatio-temporal correlations based outlier & event detection in wireless sensor networks. In *IEEE, Wireless Communications and Networking Conference (WCNC)*, pages 2048–2053, Paris, France. IEEE.
- Shih, K., Wang, S., Chen, H., and Yang, P. (2008). Collect: Collaborative event detection and tracking in wireless heterogeneous sensor networks. *Computer Communications*, 31(14):3124–3136.
- Sterk, G. and Stein, A. (1997). Mapping wind-blown mass transport by modeling variability in space and time. *Soil Science Society of America Journal*, 61(1):232–239.
- Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D. (2006). Online outlier detection in sensor data using non-parametric models. In *Proceedings of the 32nd international conference on Very large data bases, VLDB '06*, pages 187–198, Seoul, Korea. VLDB Endowment.
- Süntinger, M., Obwegger, H., Schiefer, J., Limbeck, P., and Raidl, G. (2010). Trend-based similarity search in time-series data. In *Second International Conference on Advances in Databases Knowledge and Data Applications (DBKDA)*, pages 97–106, Menuires, France. IEEE.
- Thiebaut, H. and Pedder, M. (1987). Spatial objective analysis with applications in atmospheric science. *London: Academic Press, 1987*, 1.
- Torres, M. (2006). *Energy consumption in wireless sensor networks using GSP*. PhD thesis, University of Pittsburgh.
- Vuran, M. and Akan, O. (2006). Spatio-temporal characteristics of point and field sources in wireless sensor networks. In *IEEE International Conference on Communications, 2006. ICC'06.*, volume 1, pages 234–239, Istanbul, Turkey. IEEE.
- Vuran, M., Akan, Ö., and Akyildiz, I. (2004). Spatio-temporal correlation: theory and applications for wireless sensor networks. *Computer Networks*, 45(3):245–259.
- Wang, M. and Wu, Z. (2010). Spatio-temporal correlation based outlier detection algorithm in sensor network. In *The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, volume 4, pages 424–427, Singapore. IEEE.
- Wang, T. and Yu, C. (2005). Collaborative event region detection in wireless sensor networks using markov random fields. In *2nd International Symposium on Wireless Communication Systems, 2005*, pages 493–497, Siena, Italy. IEEE.

- Wang, X., Lizier, J., Obst, O., Prokopenko, M., and Wang, P. (2008). Spatiotemporal anomaly detection in gas monitoring sensor networks. In Verdone, R., editor, *Wireless Sensor Networks*, volume 4913 of *Lecture Notes in Computer Science*, pages 90–105. Springer Berlin Heidelberg.
- Webster, R. and Oliver, M. (2007). *Geostatistics for environmental scientists*. Wiley.
- Wickham, H. (2007). Reshaping data with the reshape package. <http://www.jstatsoft.org/v21/i12/>.
- Wickham, H. (2009). ggplot2: elegant graphics for data analysis. <http://had.co.nz/ggplot2/book>.
- Wickham, H. (2012). scales: Scale functions for graphics. <http://cran.r-project.org/web/packages/scales/index.html/>. ISSN 1609-3631.
- Wu, W., Cheng, X., Ding, M., Xing, K., Liu, F., and Deng, P. (2007). Localized outlying and boundary data detection in sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1145–1157.
- Yang, Z., Meratnia, N., and Havinga, P. (2008). An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2008*, pages 151–156, Sydney, Australia. IEEE.
- Yu, Y., Jia, Z., and Zhang, R. (2012). Prediction-based algorithm for event detection in wireless sensor networks. In *11th International Conference on IEEE Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1889–1894, Liverpool, United Kingdom. IEEE.
- Zahumenskỳ, I. and SHMI, J. (2004). Guidelines on quality control procedures for data from automatic weather stations. Technical report, World Meteorological Organization (WMO), Geneva, Switzerland.
- Zhang, C., Wang, C., Li, D., Zhou, X., and Gao, C. (2009). Unspecific event detection in wireless sensor networks. In *International Conference on, Communication Software and Networks, 2009. ICCSN'09.*, pages 243–246, Macau, China. IEEE.
- Zhang, Y. (2010). *Observing the unobservable : distributed online outlier detection in wireless sensor networks*. PhD thesis, University of Twente, Enschede.

-
- Zhang, Y., Hamm, N., Meratnia, N., Stein, A., van de Voort, M., and Havinga, P. (2012). Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science*, 26(8):1373–1392.
- Zhang, Y., Meratnia, N., and Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE, Communications Surveys & Tutorials*, 12(2):159–170.