GIMA
Geographical Information Management and Applications

Integrating Formal data and Volunteered Geographic Information

A case with amateur weather data and formal KNMI data

Date:	**May 2016**
Supervisor Wageningen UR:	**Qijun Jiang, MSc.**
Professor Wageningen UR:	**prof.dr.ir. Arnold Bregt**
Supervisor KNMI:	**dr. Raymond Sluiter**
Student:	**Thomas Merkus**
UU Student number:	**3372308**
UT Student number:	**s6020461**

Preface

This master thesis is written as part of the curriculum of the master Geographic Information Management and Applications. This research is the result of an individual effort, although it would not have been possible without the support and council of: dr. Raymond Sluiter, prof.dr.ir. Arnold Bregt, and Qijun Jiang, MSc. Consequently, I would like to express my great appreciation for their contributions.

Thomas Merkus,

30 May, Utrecht.

Summary

The main objective of this research was to examine the potential benefits that are associated with the integration of formal and informal data. This was studied according to the integration of amateur weather station data derived from the WOW-NL application and formal KNMI data (temperature measurements). It is known that the integration of these types of sensor data can improve the spatiotemporal resolution of formal data sets. Hence, it was examined to what extent this was true for this specific case.

However, the use of informal data requires caution since there is a considerable lack of quality control and validation standards. As a result, it was imperative to conduct substantial pre-processing and an elaborate data assessment before the WOW-NL data could be integrated with the KNMI data.

In order to determine the quality of WOW-NL data, its observations were compared with reliable temperature estimates. Accordingly, these were derived from interpolations based on formal KNMI data. However, this required the selection of the best interpolation method for temperature measurements with a 10 minute temporal resolution. An extensive inter-comparison showed that Tin Plate Splines is the best performing interpolation method in this regard.

The assessment of the WOW-NL data showed that the WOW-NL stations generally observe higher temperatures than were estimated. Furthermore, the WOW-NL data also contains a substantial amount of gross errors and outliers that had to be removed prior to the data integration. Besides that, the WOW-NL observations deviate most from the estimates during the day in summer. Station attributes that were derived from the metadata did not show notable patterns in this regard. In addition, there were quite some stations that only showed considerable deviation from the interpolations when the predicted temperatures were above approximately 20 °C. Since amateur weather stations are known to have weak radiation shields which cause them to overheat, it is likely that these patterns are the result of radiation bias.

When the gross errors and outliers were removed, the data could be integrated. This was done according to three different integration scenarios. These include: (1) threating the WOW-NL data as equal compared to the KNMI data, (2) using the WOW-NL data only as a secondary predictive variable, and (3) making corrections for solar radiation to the WOW-NL data.

The first scenario showed that the integrated data improved temperature interpolations for the Netherlands in both October and January. However, in August the integrated data did not improve the interpolations. Equally, the second integration scenario showed that the integrated data improved interpolations considerably for October and January. For August, the improvements were negligible. The third scenario showed that corrections could be made according to incoming solar radiation. However, the corrections only resulted in a marginal improvement over the original WOW-NL data for the whole study period. When the same relation was modeled for exemplary warm days in August, more substantial corrections could be made, and solar radiation had more predictive power. Finally, the integrated datasets could be used to make maps with an increased spatiotemporal resolution.

Finally, this research concludes that the integration of the WOW-NL data and KNMI data can improve the spatiotemporal resolution of meteorological data and maps. However, the extent to which this is true is highly dependable on the time of the year and the data integration method.

Table of Contents

1. Problem and its Context (Background)

On 10 February 2011, The Times published an article which opened with the following sentence: “Flash floods, snow flurries and even tornados that pass unnoticed by the Met Office's weather stations will soon be monitored by a new network of amateur weathermen (The Times, 2011)”. Subsequently, the newspaper article describes a project of the United Kingdom’s Meteorological office (Met Office), which includes the development of a web application named the ‘Weather Observation Website’ (WOW). The main purpose of the WOW application is to connect amateur weather stations to the World Wide Web so that meteorological measurements, carried out by volunteers with amateur weather stations, can be uploaded and shared with others. With this project the Met Office aims to provide a platform for amateur meteorologists as well as to gain more data and subsequently better insight regarding local weather conditions that are hard to monitor with merely official weather stations.

Since the Met Office took action in 2011, the project has received a substantial response from volunteers. Only a year later, the data from more than 400 amateur weather stations was uploaded on a structural basis (Bell et al., 2013). As a result, the project has inspired various other meteorological institutes. Four years after the Met Office launched their web application, the Royal Dutch Meteorological institute (KNMI) launched a similar one named 'WOW-NL jouw weer op de kaart’ (KNMI 2015b). The Dutch web application serves the same purpose for the KNMI and owners of an amateur weather station in the Netherlands.

With the growing use of these types of amateur weather stations, there is now a significant amount of free sensor data available that can be used for meteorological analysis. However, Williams et al., (2011) argue that due to the nature of the data it is necessary to question its quality. Since these weather stations are being operated by amateurs, there is no control or possibility to directly verify the measurements that come from them. This should be acknowledged as an important limitation, since many factors such as for example the placement of a weather station can have a significant influence on the resulting measurements. Besides that, other issues can play a big role as well. For instance, the varying quality of weather stations that amateurs use can equally cause different results. It should consequently be acknowledged that using the data from amateur weather stations as input for further analysis and applications, requires caution and substantial pre-processing in order to filter out possible corrupted measurements (Muller et al., 2015). Not doing so could potentially propagate errors towards other parts of an analysis/application and may cause incorrect conclusions or results.

Others have also acknowledged this and subsequently carried out statistical analysis, inter-comparisons and calibrations to assess the quality of similar data (Sosko & Dalyot 2015; Bell 2014; Wolters & Brandsma 2012; Muller et al., 2015). The results of these studies are promising as they conclude that ultimately, if appropriate validation and quality control methods are used, this type of Volunteered Geographic Information (VGI) results in a valuable meteorological source with a high spatiotemporal resolution. This is especially the case for areas that are monitored by only few official weather stations, since the use of additional data points can result in a more complete and accurate estimation of local weather.

Keeping the former mentioned quality issues in mind, it is necessary to question how these apply to the data derived from amateur weather stations in the WOW-NL application, and subsequently how they can be circumvented or alleviated. In this case, it is a benefit that there is also data available from official weather stations. Accordingly, this data can be used to test the data from amateur weather

stations, as the measurements from the official weather station are supposed to be validated. Ideally, the data from amateur weather stations and official weather stations can be used to complement each other instead of existing separately. Therefore, the main research question in this master thesis is as follows:

> ***"To what extent can the integration of spatial data, derived from amateur weather stations in the WOW-NL application, and formal KNMI data improve the spatiotemporal resolution of meteorological data?"***

Furthermore, this master thesis is written according to the following structure. Chapter 2 describes the research objectives, including a formulation of the main research question and the associated sub-questions. This chapter also contains a sub-section regarding the research limitations (the scope of the research). Next, chapter 3 is dedicated to the theoretical background. This includes an overview of the most important theoretical concepts that are relevant for this research. Chapter 4 describes the methodology of the research, which includes a description of the steps that were required to realize the research objectives. Chapter 5 contains the results that were derived from performing all the steps that were described in the methodology. Thereafter, chapter 6 includes the discussion regarding the research results. Subsequently, chapter 7 comprises the most important conclusions of this research. Finally, chapter 8 includes recommendations that are based on all findings of this research. Additionally, this thesis includes a list of references as well as an Appendix, which includes various supportive research materials.

2. Research Objectives

This chapter includes a description of the research objectives that are set and guide the research. Furthermore, the aim of this research is to most adequately answer the main research question, which is formulated as follows:

> ***"To what extent can the integration of spatial data, derived from amateur weather stations in the WOW-NL application, and formal KNMI data improve the spatiotemporal resolution of meteorological data?"***

The main research question describes the integration of two specific types of spatial data. However, it should be noted that they are exemplary for both formal and informal data. Hereby, formal data is defined as spatial data that is produced and validated by official organizations that are either commercial or subsidized by a government (Cinnamon 2015). On the other hand, informal data is defined as spatial data that is collected by members of the public, i.e., volunteers. The latter includes data that is collected by sensors as well as data that is collected according to human observations. Keeping this dichotomy in mind, it should be noted that more abstractly formulated, the purpose of this research is to examine the potential benefits that are associated with integrating formal and informal data.

This is done according to a case study with amateur weather stations and formal KNMI data. Amateur weather stations in the WOW-NL application usually measure: wind direction, wind speed, temperature, rainfall, air pressure, and humidity. Besides that, it should be mentioned that there are also other cases in which volunteers monitor environmental phenomena with in situ sensors. Equivalent sensors that are managed by volunteers for instance measure nitrogen dioxide (NO2) and carbon monoxide (CO) to determine air quality (Muller et al., 2015). Characteristic for this type of data is that the phenomena that they describe are often limitedly covered by formal data providers in terms of spatiotemporal resolution (Corke et al., 2010; Muller et al., 2015). As a result, the integration with informal data could potentially complement the formal data, and subsequently improve its spatiotemporal resolution.

However, as was mentioned in the introduction, the integration of formal data and informal data comes with a variety of issues. The most important issues are related to the quality of the informal data. A central theme in this research therefore consists of analysing data derived from amateur weather stations in order to determine its quality. The methods that are used in order to do this are further clarified in the chapter 4. Furthermore, the main research question is answered according to a set of sub-questions, which are formulated as follows:

- *What are the benefits and shortcomings of Volunteered Geographic Information that is derived from sensors?*

Firstly, it is a key objective to gain a clear picture of the benefits and shortcomings that are associated with VGI that is derived from sensors. Hereby, it is questioned what exactly can be gained by using this type of data as well as what the main pitfalls are. Accordingly, these are structurally described in order to gain a comprehensive overview of how these potentially apply to WOW-NL data. As a result, this illuminates the aspects of the WOW-NL data that are important for further analysis in order to determine its quality. Next, the second sub-question has a methodological nature, and is formulated as follows:

- *How can the quality of data derived from amateur weather stations be determined?*

In order to answer this sub-question, a method was developed that is capable of determining the quality of the WOW-NL data. Besides the characteristics of sensor VGI, the formal KNMI also plays an important role in this method, as it can be used as a validated source to (statistically) compare with the WOW-NL data. The eventual method and its processes that are used in this research are described in the chapter 4. Subsequently, the third sub-question is focused on the analytical part of this research, and is formulated as follows:

- *How accurate is the data that is derived from amateur weather stations in the WOW-NL application?*

The third sub-question is answered by conducting the methods that were chosen in the previous sub-question. Consequently, the quality of the data that is derived from the amateur weather stations in the WOW-NL application could be determined. This is a crucial step in the research, as negative results could potentially alleviate the positive effects of the integration of the WOW-NL data with the KNMI data. Nevertheless, the following sub-question is posed to further clarify how the integration of both data types can be done in order to improve the spatiotemporal resolution of meteorological data and maps:

- *How can the data that is derived from amateur weather stations in the WOW-NL application improve the spatiotemporal resolution of temperature data and maps produced by the KNMI?*

This sub-question could only be answered after the informal data was analysed and validated since quality cannot be determined after the data is already integrated. Besides that, it should be noted that temperature has been chosen as an example since it is a meteorological phenomenon which is known to vary significantly over small distances (especially in urban environments) (Muller et al., 2015). Consequently, improving the spatial resolution of temperature data and maps can potentially reveal relatively much local differences in temperature that are not visible on similar maps with a coarse spatial resolution. Therefore, it is chosen to make temperature maps for the Netherlands, which have a higher spatial resolution than official KNMI maps. In order to produce temperature maps with a high spatial resolution, it is also necessary to review interpolation methods since in between observation locations values are always estimated (Meng et al., 2013).

2.1. Scope

The research questions and the associated explanations have described the overall research objectives that are set in this master thesis. Concluding, the aim of this research is to explore the benefits that are associated with the integration of formal and informal data. Especially, the focus is set on the integration of sensor VGI and official sensor data. As a case study, it is chosen to use formal KNMI data, and data derived from amateur weather stations (WOW-NL). As a prime example within this case study it is chosen to improve the spatial resolution of Dutch national temperature data and maps. However, before this can be done it is necessary to analyse the informal data so that its quality can be assessed.

Besides these overall objectives, it is necessary to further demarcate this research. It should be noted that this research does not aim to develop a generic method to integrate formal and informal data. Instead, the research is primarily dedicated to examine the resultant shortcomings and benefits that

are associated with integrating formal and informal data according to one specific case study. Since there are more examples of this type of volunteered environmental sensor data, it is perceived as a topic that might be of importance to other applications. However, this research examines the integration of formal and informal data according to one example. Hence, all results will only lead to conclusions for this specific case.

Furthermore, it should be stressed that this research emphasizes on developing an integrated spatial dataset, which confirms the potential of integrating formal and informal sensor data. In order to demonstrate the benefits of this integration, geo-visualizations will be made that illustrate the improved spatiotemporal resolution of the data. Therefore, this master thesis consists of a substantial part that is dedicated to the development of a practical geographical information system (GIS) application.

Besides that, it is also important to note that this research was commissioned by the KNMI, and conducted in collaboration with GIMA. Next to this master thesis, another similar research was carried out at the same time by (Koolen 2016). His research is characterised by a more practical approach that is primarily dedicated to data assessment regarding the WOW-NL variables: temperature, precipitation, and wind. Although there is some methodological overlap, Koolen (2016) studied multiple weather situations, the urban heat island effect, and does not review interpolation methods. Furthermore, the most important differentiation between the studies consists of the emphasis on the effects of the WOW-NL and KNMI data integration, which is characteristic for this research.

3. Theoretical background

3.1. Volunteered Geographic information & sensor data

Feick & Roche (2013) argue that 'crowdsourced' spatial data or Volunteered Geographic Information are terms that are used to describe geographic information that is generated by volunteers, i.e., by others than commercial or governmental organizations. According to them, the term VGI is widely applicable to many different types of data that can be used for many different purposes. They state that:

> "[…] VGI ranges from data that are experiential and largely personal in nature (e.g. geotagged vacation photos) through passively contributed information concerning personal activity spaces (e.g. credit card transactions, cellular phone tracking) to what might be considered quasi-scientific data (e.g. locations of animal sightings, amateur weather station readings) (Feick & Roche 2013, p.22)."

In order to comprehend the wide range of VGI, it is necessary to briefly touch upon its origin. The term VGI was first introduced by Goodchild (2007), as he argued that: "VGI is a special case of the more general Web phenomenon named user-generated content (Goodchild 2007, p.212)". He states that, initially, the World Wide Web was characterized by an environment in which users were mainly consuming content that was predefined by the developers of websites. As the World Wide Web developed during the past decade, it became more user-centred, interoperable and users increasingly started to contribute to websites by sharing their own data; which is referred to as 'user-generated content'. As a result, this trend initiated a new collective way of using the World Wide Web, which is often referred to as 'the Web 2.0.'.

Accordingly, Goodchild (2007) argues that the development of the Web 2.0. has been essential for the arrival of VGI as it created an environment that made it is much easier to share geographic information. Besides that, the arrival of relatively cheap sensors and GPS equipped devices (e.g., smartphones, cameras) has also contributed to the growth of VGI. This is due to the fact that this made it much easier for the public to gather spatially referenced data. This is also what sets VGI apart from other user generated content, as VGI has to refer to a place on earth. Hence, a relatively diverse range of data can be considered as VGI.

Besides its many different forms, VGI is also used for a wide range of different purposes. Some VGI applications are dedicated freely map the world, while others are for instance focused on improving disaster management or environmental monitoring (Zook et al., 2010). What most types of VGI applications have in common is that they aim to produce data for which there is no viable alternative. This can for instance be caused by the absence of similar data, or the high prices that formal data providers ask. Hence, it should be acknowledged that, despite the considerable variation in the type and purpose of VGI, altogether its power lies in illuminating local activities in various places that go unnoticed by the world's media (Goodchild, 2007). Besides that, VGI is only dependent on volunteers, so the potential of VGI is inherently linked to the large amount of volunteers that can contribute as well as to the amount of data that they can produce.

Conversely, there are also some important limitations that apply to most types of VGI. Firstly, VGI is often characterized by a lack or absence of metadata. This is due to the fact that many VGI applications do not make it mandatory to include metadata (Flanagin & Metzger 2008). Furthermore, the quality of VGI is frequently disputed, as there are no formal methods of control in most cases. As a result, VGI can for instance lack in completeness, accuracy or other quality aspects that would normally be

inspected with the production of formal data. Also, VGI is known to have a subjective spatial coverage, since volunteers tend to prefer generating VGI for popular places. As a result, urban areas are commonly better covered than remote places. Nevertheless, it should be noted that the most important benefits and shortcoming that are associated with VGI tend to be case and type specific.

Accordingly, it is also important to take the nature of VGI into account. While many types of VGI are acquired by means of human observation (e.g., animal spotting, textual description, perceptions of safety), others come from sensors (e.g., camera, GPS, amateur weather stations) (Haklay 2013). It goes without saying that data derived from sensors should be acknowledged as more objective than human observations as sensors cannot be biased in ways that humans tend to be. However, it should be stressed that sensors do come with their own range of issues.

3.1.1. Sensor data

In general, there are many different examples of sensor data that is crowdsourced by volunteers. Muller et al., (2015) differentiate between three different types of crowdsourced sensor data, namely: active, passive, and semi-passive (figure 1). Active crowdsourcing requires humans to constantly interact with applications in order to generate output. This means that volunteers have to interact with a device on a constant basis in order to generate data. With passive crowdsourcing, no human interaction is required in order to collect and subsequently upload data.

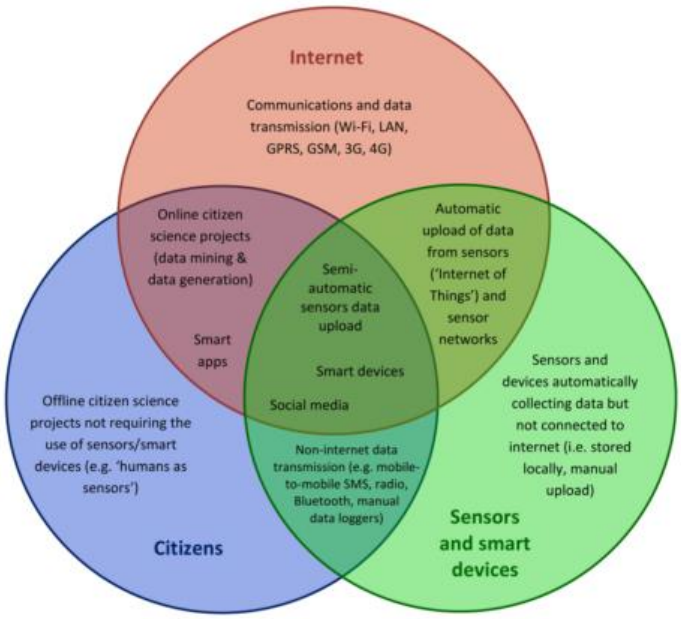


Figure 1: Venn diagram showing the interaction of animate and inanimate crowdsourcing components, including active and passive techniques, (Muller et al., 2015, p.37).

In this case volunteers merely act as regulators, since they only have to manage the sensor device by its installation and maintenance. Automated weather stations for instance fall into this category since they automatically push the measurements to a server or another device via WIFI, while (in theory) they only have to be installed once. Besides that, semi-passive crowdsourcing falls somewhere in between, and only requires human interaction when data needs to be published to a server. Subsequently, it is also of considerable importance whether a sensor is 'static' or 'mobile'. Static sensors are in situ sensors that do not move, while mobile sensors are used to measure phenomena

while moving across space (multiple locations). Mobile sensors are usually not mobile by themselves as they are rather mounted to a mobile platform (Castell et al., 2014).

Although the issues that arise with using sensor VGI are heavily dependent on its specific type and user case, some common issues can be recognized as well (Muller et al., 2015). Firstly, the maintenance of VGI sensors can become an issue since its responsibility entirely lies with the volunteers that own them. Accordingly, a lack of maintenance can cause for malfunction and lead to corrupted results. Furthermore, calibration errors can equally cause for artificial biased results, as volunteers are often not willing or capable to (re)calibrate their sensors. Thirdly, the continuity of the data output can also be a factor that can become a limitation since there are no guarantees that volunteers will consistently manage a sensor and forever generate data. Of course, this is also true for formal data providers, although they should be considered as more reliable. Besides that, issues such as signal loss can be beyond the control of a volunteer and equally disturb data continuity (Chapman et al., 2014; Muller et al., 2015).

3.1.2. Amateur weather stations

Although the common issues that are associated with crowdsourced sensor data are relevant, it is even more important to consider case specific limitations. Bell (2014) also acknowledges this, and has resultantly researched which factors can potentially cause issues for data that is derived from amateur weather stations. According to (Bell 2014, pp.40–47) the following factors can cause uncertainty: calibration issues, design flaws, communication and software errors, metadata issues, and representativity errors. Subsequently, he describes them as follows:

Calibration issues

Amateur weather stations can be wrongly calibrated, as they can be biased from before installation or they can drift gradually over time. Calibration is an important process since this makes measurements accurate, consistent and comparable with measurements from other devices. This part is however frequently carried out wrong or not at all, as it is often too complicated or expensive for volunteers. This is an important difference with professional weather stations, as their sensors are recalibrated on a structural basis, making sure that they stay accurate.

Design flaws

The design of amateur weather stations often compromises meteorological standards and resultantly makes them susceptible to inaccurate measurements. These design flaws are usually cost-driven as the manufacturers aim to keep the amateur weather stations affordable for volunteers. Besides that, they also try to keep the weather stations user-friendly and aesthetically appealing. These compromises resultantly degenerate the quality of the weather stations. One of the most common design flaws includes a weakly manufactured radiation shield, which results in overheating. Professional weather stations are not bound by these market-driven compromises and primarily prioritize accurate measurements. It is therefore of considerable relevance to know which weather station generated what data so these design flaws can be taken into account.

Communication and software errors

The communication and software related issues can cause errors of relatively large proportion, as well as extensive gaps in the data. After the measurements are converted into an electrical signal they have to be transported from the weather station to a (online) database. Any errors that might occur in this process should be acknowledged as communication and software errors. Since amateur weather stations often transport data via WIFI, these errors can be caused by issues as simple as connection loss/disturbance or software errors on the receiving end. Mostly, these errors result in wrongly

assigned time steps or data gaps. Accordingly, it should be relatively easy to filter out these errors with appropriate quality control methods.

Metadata issues

As is often the case with spatial data, the lack of metadata makes it difficult to interpret its usability and quality; this is also evident for data derived from amateur weather stations. Issues in this regard arise because volunteers are not obliged to follow instructions for providing metadata. Examination of WOW data as well as data from Weather Underground makes it clear that, although there is relatively much metadata generated, some aspects remain neglected as for instance a substantial amount of volunteers do not include the elevation of their weather station. Accordingly, this information is crucial for an accurate interpretation of their measurements since elevation is directly related to some of the meteorological phenomena that amateur weather stations measure. Professional weather stations and meteorologist are obliged to provide metadata for all their measurements. Since their weather stations are already of a higher quality it should be acknowledged that for volunteers it is even more important to provide extensive metadata.

Representativity error

A representativity error occurs when a weather station samples a meteorological aspect at a spatial scale, which is different from the scale that is used in a specific application. This means for example that when a thermometer measures temperature in a small garden in a built environment, those temperatures are not representative for the surrounding area since conditions are fairly different in the garden as opposed to the built-up environment. It is therefore very important to know where amateur weather stations are installed, since this determines for what areas they monitor representative data. This is a complex responsibility for volunteers, since often they install weather stations in their garden while living in an urban environment. Even in their garden, temperatures can vary significantly if there is variation in terms of vegetation, exposure, etc. Professional weather stations are generally well placed in regard to their aimed scale of representation, as these issues are taken into account. Furthermore, these types of errors can be worse than other errors and can exist despite off a perfectly working weather station. It is therefore crucial to take the placing and representativity into account when examining data derived from amateur weather stations.

Above, the most important benefits and shortcomings regarding VGI that is derived from sensors have been described. Although there is a strong conviction that VGI has much potential, it has become clear that some sort of quality control is imperative. Since data derived from amateur weather stations should be viewed as a form of VGI, there is an evident need for quality control in this case. Accordingly, prior research has indicated that an important element that is required for quality control includes a good estimation of the weather at the sites of amateur weather stations (Bell 2014). These estimations can be gained by making interpolations based on the measurements of official weather stations. As a result, it necessary to further elaborate on the most important aspects of interpolation methods, this is done in the next paragraph.

3.2. Spatial interpolation

Interpolation is mathematic approach that is often used in geography to predict values of continuous spatial phenomena at unmeasured locations according to observed values at other locations. In most cases it is not feasible to measure continuous spatial phenomena (e.g., snow depth, temperature, rainfall) for all places in a given area, hence making sensible estimates is essential. Theoretically, interpolation is based on Tobler's first law of geography, which states that: "Everything is related to everything else, but near things are more related than distant things (Tobler as cited in Meng et al.,

2013, p.28)". Accordingly, interpolation methods assume that the values of continuous spatial phenomena are to certain degree related to each other, dependent on their location. As a result, interpolation methods are able to make relatively accurate value predictions for complete continuous surfaces, based on measured values at sample sites, combined with the distance to unmeasured sites.

Evidently, not all continuous spatial phenomena are distributed across space in a similar fashion. Some are for instance characterized by a higher degree of spatial autocorrelation than others. Hereby, spatial autocorrelation refers to: "*a measure of the degree to which a set of spatial features and their associated data values tend to be clustered together in space (positive spatial autocorrelation) or dispersed (negative spatial autocorrelation)* (ESRI 2015c)". Besides that, other factors than distance can have a significant influence on the values of continuous spatial phenomena as well (e.g., elevation on temperature, surface type on albedo effect). Hence, there are many different interpolation methods, which can take distance, spatial distributions, and ancillary data into account.

Furthermore, interpolation methods can provide estimates for any given location, however for computing and visualisation purposes interpolations are mostly conducted for raster grids. This results in raster datasets that have one predicted value that is valid for a whole raster cell, which represents an area on earth dependent on its selected cell size and location.

As interpolation methods yield varying results dependent on the spatial phenomenon they aim to predict, it is important to use the method that gives the most accurate output. This means that the predicted values should resemble the values that can be observed in the real world as close as possible. The accuracy of interpolations can be verified according to various methods (e.g., cross-validation, data splitting). However, before these methods will be elaborated, it is necessary to review the most important interpolation methods that are relevant for this research.

3.2.1. Interpolation methods

Firstly, it should be noted that besides the interpolation methods that will be discussed in this research, there are many others that are used in a wide range of different disciplines. However, the focus in this research lies on the interpolation methods that are used to interpolate climatological phenomena, and especially the methods that can be used to interpolate temperature.

According to Li & Heap (2008), interpolation methods can be sorted into three main categories. These are 'geostatistical' methods, 'non-geostatistical' methods, and 'combined' methods. Non-geostatistical methods, or deterministic methods, only consider the values and the geometric properties of sample observations in order to predict values at unknown locations. Methods that fall into this category for instance include 'Inverse Distance Weighting' (IDW), 'Thin plate splines' (TPS), and 'Nearest Neighbor' (NN).

Conversely, geostatistical methods make use of both mathematical and statistical properties of sample observations in order to predict values at unknown locations. Accordingly, geostatistical methods measure the spatial autocorrelation of sample observations and subsequently include these in the prediction for unknown values (Dobesch et al., 2007; Dyras et al., 2005; Joly et al., 2011; ESRI 2015d). Moreover, geostatistical methods also give a probabilistic estimate of the quality of the predictions. The geostatistical interpolation methods are all based on the works of Danie Gerhardus Krige, hence all the varying methods are named as a specific type of 'kriging' interpolation.

Finally, combined methods are usually developed for specific purposes and integrate aspects from both geostatistical methods and non-geostatistical methods (Li & Heap 2008; Sluiter 2012). Examples include: 'Trend Surface Analysis Combined with kriging' and 'Lapse Rate Combined with kriging'.

3.2.2. Important features

Besides the former described categorization, interpolation methods have some fundamental features that are of considerable importance as well. Accordingly, Dyras et al., (2005) state that interpolation methods can either be 'global' or 'local'. This refers to whether or not an interpolation method takes all sample observations into account for the prediction of an unknown value. In case all values are taken into account, it should be considered as a global method. Conversely, if only a few sample observations that are within a specific area around an unknown value are considered, the interpolation method should be acknowledged as local.

Furthermore, it should be noted that there is also a relevant difference between exact interpolation methods and inexact interpolation methods. Exact interpolation methods result in continuous surfaces, in which values at sample sites exactly match the values of the sample observations. Conversely, inexact interpolation methods do not reproduce the similar values as the original sample observations. Inexact interpolation methods do not reproduce these values in order to improve the smoothness of the interpolated surface (Lam 1983; Sluiter 2009).

Another important distinction between interpolations methods is based on whether they produce gradual or abrupt surfaces. The smoothness of the surface is determined by how the weights are assigned to the sample observations. While some methods result in predictions with spikes, others result in relatively smooth surfaces (Li & Heap 2014).

Furthermore, it is also important to distinguish between univariate and multivariate interpolation methods. Hereby, univariate methods refer to interpolation methods that only take samples of the primary variable into account. On the other hand, multivariate interpolation methods are also capable of using secondary variables for their predictions. Accordingly, these methods are capable of dealing with ancillary data. Often, these multivariate interpolations can lead to more accurate results since in reality more aspects than distance and spatial autocorrelation influence the distribution of spatial phenomena (Li & Heap 2014; Li & Heap 2008).

3.2.3. Non-geostatistical methods

Nearest neighbor

Firstly, one of the notable non-geostatistical methods includes the Nearest Neighbor method (NN). The NN method predicts the values at unknown locations by giving them a similar value to the nearest neighboring sample observation. This technique is also known as Voronoi triangulation, as it divides the interpolation surface in Voronoi triangles based on the distance between the sample observations. Consequently, for each observation there is a triangle polygon with the observation site in the center. For all other unknown points that are located in this polygon, it is evident that they are the closest to the observation sample point. As a result, the NN method creates unrealistic, abrupt surfaces, and is only sporadically used in meteorology (Sluiter 2009; Li & Heap 2008).

Inverse Distance Weighting

Subsequently, the Inverse Distance Weighting (IDW) is another notable, local interpolation method although it takes more than one neighboring sample observation into account. Accordingly, IDW takes all sample observations into account within a certain radius around the unknown value that has to be predicted, and gives them weights according to their distance from that location. This way, IDW assigns bigger weights to sampled observations that are closer to the unknown value, and the weights diminish as a function of distance (integrating Tobler's first Law of geography). These weights are proportional to the inverse distance raised to the power of q (while $q > 1$). Although, the standard

value for q = 2, this not a mandatory value; testing different values may result in a more appropriate value for q. The results from the IDW can be assessed with cross-validation. Finally, it should be noted that IDW is a method that is often used in meteorology, for instance for temperature interpolations (Hofstra et al., 2008; Li & Heap 2008; ESRI 2015a).

Splines

Spline interpolation uses multiple polynomial functions to fit a trend line through all sample observations; hence it's a global exact method. However, each polynomial is fitted locally, exactly through a set of points, and subsequently connected to other polynomials in order to generate one smooth line (Li & Heap 2008). Furthermore, splines are often used in meteorology and are frequently used for the interpolation of temperature. However, it should be noted that it is more appropriate for interpolating at low temporal resolution (monthly, yearly averages). To measure the quality of the interpolation it also justified using a cross-validation.

Linear regression

The last non-geostatistical method that is worth discussing is the linear regression method (LM). In essence this model should be acknowledged as a regular linear regression model, hence the model assumes that the data is normally distributed, independent, and has a stable variance. Furthermore, linear regression expresses the relation between a predicted primary variable and the explanatory secondary variable(s). It does so by fitting a straight (linear) line through all the sample observations that results in the smallest sum of squared residuals. Consequently, it should be noted that LM is a global, inexact interpolation method. In meteorology, LM is often used, however in most cases it is used in combination with other methods (Sluiter 2009; Dobesch et al., 2007; Li & Heap 2008).

3.2.4. Geostatistical methods

Geostatistical interpolation methods are inspired by the assumption that spatial autocorrelation is an important factor for the prediction of continuous spatial phenomena. As a result, the kriging interpolation methods are based on the assumption that the variation of spatial phenomena is too complex to be quantified according to deterministic or mathematical methods only (Oliver & Webster 2014; Sluiter 2009). Consequently, kriging interpolation methods treat the variation as if it is random, and therefore describe the spatial variation with a stochastic surface, including an attribute which is named a 'regionalized' variable (Sluiter 2009). Accordingly, the kriging interpolation methods assume that the value of a random variable Z at (x) is given by:

$$Z(x) = m(x) + \varepsilon'(x) + \varepsilon''$$

Where: m(x) = a deterministic function describing a structural component of Z at x.
ε'(x) = a random spatially correlated component.
ε''(x) = a residual non-spatially correlated term, or noise (Nugget variance).

Formula 1: kriging principles, (Sluiter 2009, p.13).

If the variation is homogenous and the structural effects are taken care of, the semivariance y (h) can be estimated with the following formula:

$$\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}\{z(x_i) - z(x_i + h)\}^2$$

Where: n = number of pairs of sample points of observations of the values of attribute z separated by distance h.

Formula 2: Estimating the semivariance, (Sluiter 2009, p.13).

Subsequently, plotting y(h) against h results in a variogram (figure 2), and gives an quantitative indication for the spatial arrangement (spatial autocorrelation) of the sampled observations. Variograms are of considerable importance for geostatistical interpolation methods, since they provide estimations for the most appropriate weights for sample observations.

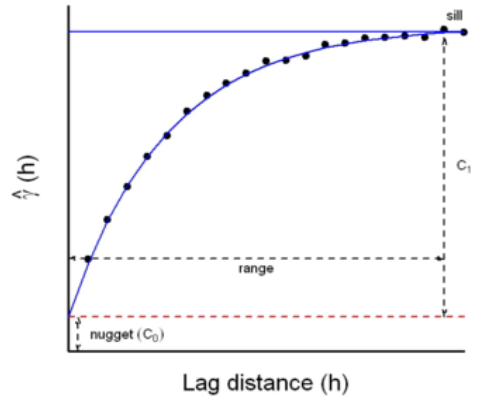


Figure 2: Example of a variogram, (Li & Heap 2008, p.12).

Furthermore, there are some important features of the variogram that require some additional explanation. As can be observed in figure 2, the model levels out at a certain amount of spatial lag. Subsequently, the first point at which this happens is called the 'range' (ESRI 2015b; Oliver & Webster 2014). All sample observations within a distance smaller than the range are characterized by certain degree spatial autocorrelation, as opposed to all sample observations that are further apart. Furthermore, the value on the y-axis at which the range is reached is called the 'sill'. Finally, the value at which the semivariogram model intercepts the y-axis is named 'the nugget'. The nugget might seem strange since at zero spatial lag there should be no variation in theory. However, the nugget represents the micro-scale variation (Dyras et al., 2005). Altogether, the nugget effect can be caused by measurement errors, or by variation at scales that are smaller than the measurement interval. This is for instance often the case with minerals, since at micro scale they can still be characterized by some variation as they are usually not distributed equally over a small patch but rather concentrated at certain points.

Ordinary kriging

The ordinary kriging (OK) interpolation method is quite similar to IDW, as it is a linear combination of measured sample observations. However, with OK the weights that are assigned to the different sample observations is not directly determined by distance, but by the spatial correlation as is described by the variogram (Dyras et al., 2005). This way the spatial autocorrelation is taken into account for the interpolation. In meteorology, the OK method is not often used stand-alone, but rather as part of other methods (e.g., Residual kriging, Indicator Kriging) (Sluiter 2009).

Universal kriging

The Universal kriging (UK) interpolation method, or kriging with external drift, assumes that the mean is not constant and that there is an overriding trend in the data which can be modeled by a deterministic function (ESRI 2015b; Sluiter 2009). As a result, this makes it possible to use ancillary data, although this should only be done if there is a scientific justification. In environmental science there are many of these relationships (e.g., height and rainfall), which can be taken into consideration with UK. Consequently, the use of ancillary data can improve results; hence UK is often used for meteorological interpolations.

Cokriging

The Cokriging (CK) interpolation method is a multivariate variant of the OK method and incorporates a multivariate variogram or covariance model with multivariate data (Dyras et al., 2005). The value predictions with CK are based on a linear sum of the explanatory co-variables. It should be mentioned that the model can become relatively complex if many co-variables are used (Sluiter 2009). Nevertheless, the CK interpolation method is often used in meteorology, and can sometimes lead to satisfactory results, especially if spatial correlation between variables is high (Sluiter 2009).

Residual kriging

The Residual kriging (RK) interpolation method firstly uses a regression model to predict values for variable that has to be interpolated. Secondly, the residuals have to be calculated, based on the comparison with the sample observations. Next, the residuals are interpolated for the whole model with the OK interpolation method. Thereafter, the predicted values and the interpolated residual values are summed to obtain the final prediction (Li & Heap 2008; Dyras et al., 2005). RK can be used for meteorological phenomena, and is for instance used to interpolate monthly temperatures averages in the United States (Wu & Li 2013).

Indicator kriging

The Indicator Kriging (IK) interpolation method is dedicated to interpolating categorical variables. Accordingly, the IK method is not used to get precise predictions, but more to gain insight regarding the uncertainty of spatial variables (Li & Heap 2008). It can for instance be used for the interpolation of climatological phenomena such as the occurrence of rainfall (Sluiter 2009).

3.2.5. Combined methods

Based on the geostatistical interpolation methods, non-geostatistical methods and other statistical approaches, many combined interpolation methods are developed in order to interpolate unknown values for continuous spatial phenomena. While there are many more, there are only two examples described in order to give an indication of how combined methods are constructed.

One of the combined interpolation methods which haves multiple variations, combines regression analysis with kriging and is known as Regression kriging. One of its notable variations is known as “kriging combined with (linear) regression” and conducts regression analysis and subsequently uses OK on the resultant values. Accordingly, this method makes use of geostatistical interpolation techniques as well as non-geostatistical techniques.

Besides that, there are many other possible combinations. There are for instance also combined interpolation methods that merely use non-geostatistical methods such as the Linear regression combined with Inverse Distance Weighting (Li & Heap 2014; Li & Heap 2008). Altogether, it should be noted that these combined methods are usually developed for specific applications and can in some cases lead to improved interpolation results.

3.2.6. Quality indicators

Due to the wide range of different interpolation methods, there is a need for indicators that describe their performance, i.e., their accuracy or precision. Li & Heap (2011) have also stressed this concern, and subsequently conducted an elaborative comparison between the various performance indicators that are commonly used.

According to Li & Heap (2011), the quality of inexact interpolation methods can be assessed by examining the difference between the interpolated values and the observed values at sample sites. Based on either the absolute differences or the squared differences, various performance indicators can be calculated (Li & Heap 2011). The Mean error (ME) for instance compares the model predicted- and the observed means, and is mainly used in order to describe average model bias. This way, it can be assessed whether a model tends to averagely over-predict or conversely, under-predict the interpolated values. However, this measure should be interpreted with caution as negative values and positive values counteract each other, which can result in a lower ME than the actual errors (Li & Heap 2011; Willmott & Matsuura 2005). Furthermore, the Mean absolute error (MAE) is a performance indicator that considers absolute errors. Accordingly, the MAE can be calculated by summing all the absolute errors and subsequently dividing them by the number of sample observations. The MAE is a performance indicator that gives the mean of the absolute errors, and is therefore susceptible to infrequent big errors.

To overcome the issue of dealing with infrequent big errors, the Mean square error (MSE) and the Root mean square error (RMSE) can be evaluated. The MSE is calculated by firstly summing the individual squared errors, and secondly by dividing them by the number of sample observations. In order to calculate the RMSE, the square root of the MSE has to be taken. Accordingly, the RMSE and the MAE are measures that describe the size of the average errors, however the RMSE gives relatively large weights to outliers. This is also the case with the MSE as both measures consider the squared difference between the predicted values and the observed values. As a result, it should be noted that although the RMSE and the MAE are fairly similar, the MAE is less sensitive for large errors. It is therefore advised to check both performance indicators to examine the influence of potential outliers. Nevertheless, it should be realized that these measures merely give an estimation of the size of the average error but they do not give any explanation for what causes the errors, or about the average difference between the errors (Li & Heap 2008).

Besides the former mentioned performance indicators, there are many others (e.g., Willmott's D and the Averaged standard error), however these are all only slightly different and will not all be discussed due to their limited relevance for this research. Finally, it should be stressed that examining the former mentioned performance indicators gives a sufficient overview of how accurate an inexact interpolation is.

Cross-validation

Furthermore, it should be noted that there are also performance indicators that can be used to measure the success of both exact and inexact interpolation methods. The most notable performance indicators can be obtained by conducting a cross-validation or the 'leaving one out' method. Accordingly, the method is performed by leaving a sample observation out, by pretending it does not exist. Instead, the value for the sample observation site is interpolated, based on all other sample observations. Thereafter, the difference (residual) with the real sample observation can be calculated by subtracting the new interpolated value. This procedure has to be repeated for all sample observation in order to get the residuals for all sample observations. Altogether, the evaluation of the

cross-validation can be obtained by getting the RMSE or the MSE of the residuals (Tomczak 1998). The cross-validation gives a good indication to what extent the interpolation model is capable of predicting unknown values, without the use of additional samples. Consequently, it is a method that is well suited for exact interpolation methods as they do not have differences between predicted and observed values at sample observation sites.

Data splitting

Besides cross-validation, another method to derive performance indicators for spatial interpolations is by means of 'data splitting'. Accordingly, data splitting is conducted by dividing the sample observation dataset into a training set and a validation set. Firstly, the interpolations are carried out for the training set. Thereafter, the sample observations that are part of the validation set are compared to the interpolated values. Performance indicators can be derived from data splitting by calculating the MSE and the RMSE from the residuals. However, it should be mentioned that for data splitting there should be sufficient sample observations available since half of them are not used anymore for the interpolation (Williams et al., 2011; Sluiter 2009).

4. Methodology

This chapter elaborates on the methodological steps that were conducted in order to answer the research questions that were posed in chapter 2. Furthermore, figure 3 gives a global methodological overview of the workflow regarding this research. Firstly, the theoretical concepts were examined and extensively reviewed. Secondly, the informal data was assessed for its quality, whereafter the informal data could be corrected for possible corrupted measurements. After these three steps, the data integration could take place in order to improve its spatiotemporal resolution. This was done according to three different integration scenarios. Finally, the integrated data was used as input to make temperature maps that show the improved resolution of the integrated data. However, it should be noted that these steps consist of more detailed methodological choices and complex processes that are extensively described in the following paragraphs.

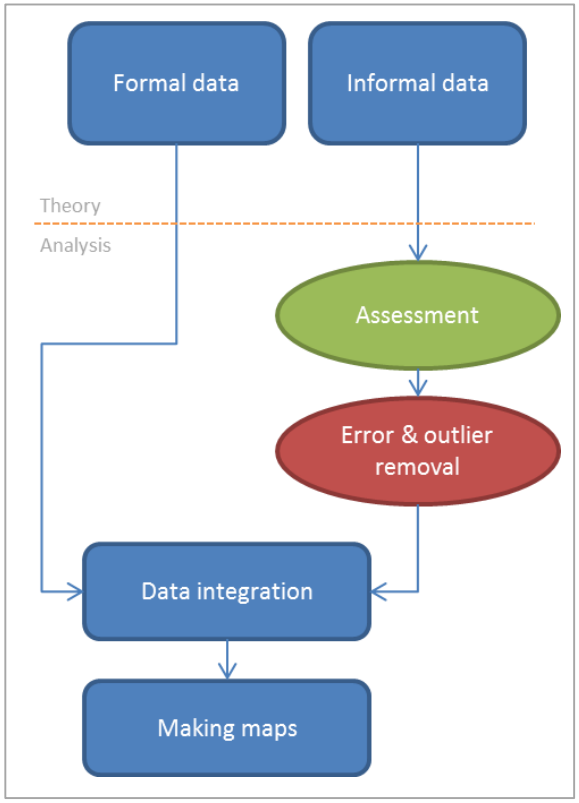


Figure 3: Conceptual overview of workflow.

4.1. Informal data assessment

The assessment of the informal data is one of the most important parts of this research. In this step the objective was to examine the accuracy of the data that is derived from amateur weather stations in the WOW-NL application. Hereby, it was important to test which part of the data is accurate and which part is biased. It should however be noted that this task is not clear-cut, and assessments of the data will always include a degree of uncertainty. Nevertheless, Bell (2014) has carried out similar research and developed a method that provides useful guidelines.

Firstly, Bell (2014) states that in order to test whether any given amateur weather station is biased, it is necessary to have an reliable independent estimation of the weather at the site of the amateur weather station (figure 4). Accordingly, these estimations can be used to compare with the

measurements that have been carried out by volunteers. As a result, these comparisons give insights into the quality of the amateur measurements since the size of the differences tells something about the plausibility of those amateur measurements being correct.

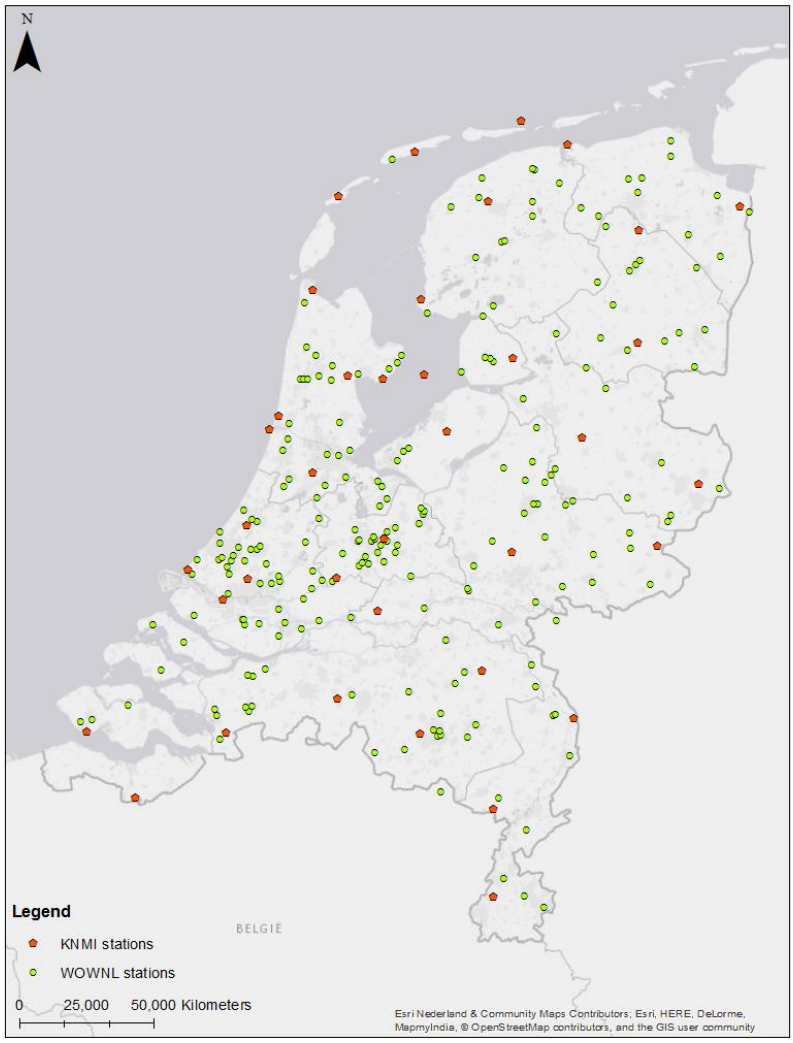


Figure 4: Spatial distribution of Automatic weather stations (KNMI stations) and WOW-NL stations.

Consequently, this step of the research consists of two parts. The first objective is to find the interpolation method that gives the most accurate predictions of temperature values at the sites of amateur weather stations, based only on automatic weather station (AWS) measurements (official KNMI stations). Next, the second objective is to extensively compare the predictions that are based on professional measurements with the WOW-NL data.

4.1.1. Finding the best interpolation method

The first step in this part of the research is to find an interpolation method that gives the 'best' estimates of temperature at the sites of amateur weather stations. This is done according to measurements derived from automatic weather stations (figure 4). The method that is used to find the most appropriate interpolation method is explained according to the three questions: *How? What? & When?*

How?

In this research it is chosen to find the 'best' interpolation method for temperature in the Netherlands at a temporal resolution of 10 minutes. This is done by comparing four of the most promising interpolation methods for three different months in different seasons. The results of all these interpolations are tested for quality according to a cross-validation as was described in the theoretical background. This means that for every interpolation, the leave-one-out cross-validation was conducted and quality indicators were calculated. Accordingly, the quality indicators that are evaluated include the RMSE and the ME.

In order to determine which method is the best, quality indicators are not the only criteria are evaluated. Instead, the three criteria 'stable ', 'replicable', and 'high quality', as described in table 1, together decide which interpolation model should be used for the second step of the data assessment procedure. These criteria were selected during an extensive discussion with experts (dr. Raymond Sluiter, prof.dr.ir. Arnold Bregt, and Qijun Jiang).

Selection criteria:	Description:
Stable	The RMSE of the model (tested with cross-validation) should be relatively stable through time & the ME should not structurally over- or under predict.
Replicable	The method should not become too complicated, and easy for others to understand and to replicate.
High quality	The quality should not be significantly lower than other methods in terms of quality (RMSE), and preferably the best.

Table 1: Selection criteria for the most appropriate interpolation method.

Firstly, a stable model is essential since thousands of interpolations are compared with amateur weather data. If the model performs well on average but has relatively much and big outliers, those comparisons will become irrelevant.

Secondly, it is chosen to also prefer less complex models to complex models. The reason for that can best be underpinned by one of Einstein's most famous quotes: "Everything should be made as simple as possible, but not simpler (Einstein 2010, p.475)". Accordingly, this quote compactly summarizes the theory of Occam's razor, which prescribes that if there are two equally performing, competing hypotheses which are trying to predict the same thing, the less complex one is better (Blumer et al., 1990).

Finally, the last selection criterion is purely based on the RMSE and ME quality indicators. Evidently, it makes sense to use a model that outperforms competing models. This is especially true if the differences between them are considerable; in that case it is preferred to select the one with the lowest RMSE and ME.

What?

As was described in the theoretical background, many interpolation methods exist that can be used to interpolate air temperature in the Netherlands. Besides that, it has also been stressed that finding the most appropriate method dependents on the type of spatial phenomenon that has to be predicted and the available data. As a result, it has been chosen to test four different methods that are commonly used to interpolate temperature in the Netherlands, namely: IDW, OK, UK, and TPS (Dirksen 2015; Salet 2009; Sluiter 2012). As is illustrated in figure 5, these methods result in very different maps representing very different spatial distributions. Besides the different results, these techniques also have different ways of making the calculations. As a result, this also means that different parameters and methodological steps had to be chosen in order to come to the results. Accordingly, these are described in the following paragraphs.

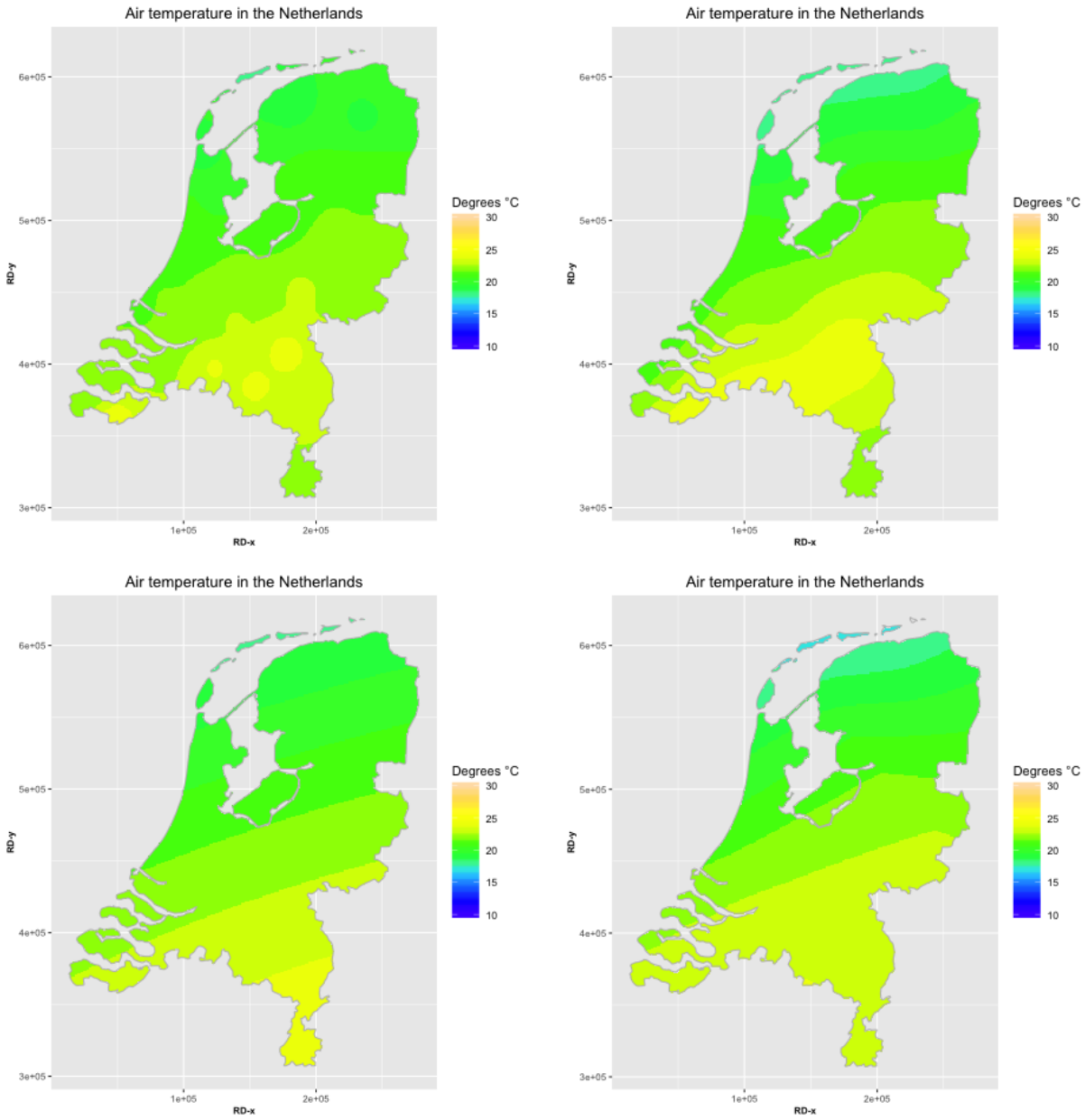


Figure 5: Air temperature on 8-8-2015 17:50, made with different interpolation methods based on AWS data. From top left to bottom right: IDW, TPS, OK, and UK.

Inverse Distance Weighting

As was described in the theoretical background, the IDW interpolation method is one of the most straightforward models that have been tested. Accordingly, only a few parameters have to be set and calculations are relatively easy. In this case the standard power function (2) has been used, since small scale testing showed that tweaking the power function shows only little improvements and fluctuates per interpolation. Furthermore, the amount of neighboring measurements that have to be taken into account has been set to the max, making it a global interpolation method. This is done since the area only covers the Netherlands, which is not too large and relatively uniform. Besides that, the automatic weather stations do not provide a big sample as most of the times circa 30-35 stations provide measurements. Finally, it should be noted that these interpolations were done using the gstat package in the R programming language, hence making it possible to automate it for the entire study period (Pebesma & Graeler 2016).

Ordinary Kriging

The OK interpolation method is one of the more complicated methods that have been tested in this research. Overall, the OK method is much like IDW, although instead of distance OK uses spatial autocorrelation to determine weights of surrounding observations. Accordingly, the OK method requires for every interpolation that a sample variogram has to be made, and a model has to be fitted. Hence, the OK method requires some additional elaboration.

Firstly, it should be noted that these procedures had to be automated, as they needed to be repeated thousands of times. Consequently, the Automap package for the R programming language was used (Hiemstra 2015). This package was built in order to make it possible to test multiple variograms models as well as multiple nugget-, sill-, and range parameters for all interpolations. The objective is to test multiple models and parameters, and to select the best configuration each time.

The parameters that have to be set include: the sill, the range, the nugget, and the fitted models. Firstly, the sill is estimated by taking the mean of the maximal value and the median value of the semivariance. Furthermore, the range is chosen by multiplying the diagonal of the bounding box that is associated with the dataset. The nugget is set to the minimal semivariance value. It should be noted that these procedures are all standard practice in the Automap package and turn out to give optimal results as is illustrated later (Hiemstra 2015).

Besides these parameters, the variogram models that have to be tested need to be specified as well. In this case the most common variogram models are tested, these include: the Exponential, the Gaussian, the Spherical, and the Matern, M. Stein's parameterization (figure 6). The Matern, M. Stein's parameterization does not have a standard form hence it is not illustrated. In order to select the most appropriate model, the difference between the sample variogram and the fitted model is evaluated. Accordingly, the model that has the smallest sum of the squared residuals is chosen. Once the most appropriate variogram model is chosen, it can be used to determine the weights of the surrounding observations, as was the case with IDW. Thereafter, the predictions can be made and be assessed with a cross-validation. Finally, it should be noted that this entire selection procedure had to be done for every time step; hence using a programming tool is the only viable solution.

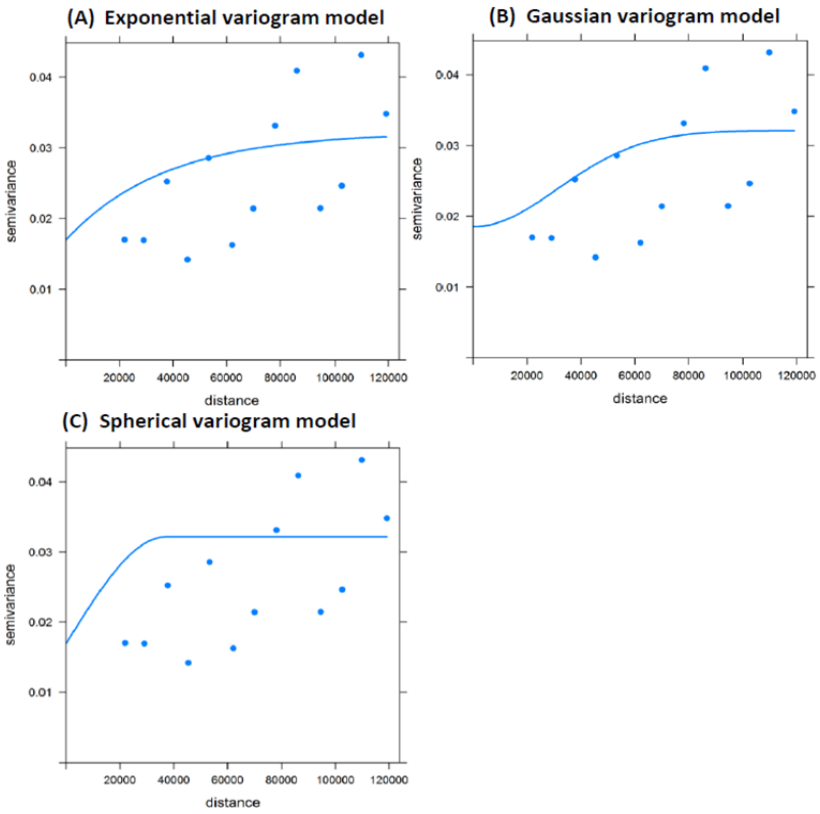


Figure 6: Examples of variogram models, source: Dirksen 2015, *p.36.*

Universal Kriging

Closely related to the OK interpolation method is the UK interpolation method. As was stated in the theoretic background, the UK interpolation method can be used when there is an overriding trend in the data, which can be modeled by a deterministic function. For air temperature in the Netherlands, previous research has revealed that the distance to shore can be used as such explanatory variable (Dirksen 2015; Salet 2009; Sluiter 2012). Accordingly, the distance to the shore was digitalized in a GIS, and used as an explanatory variable in this research (figure 7). It should however be noted, that the logarithmic distance to shore shows the best results, hence it is also used in this research (Salet 2009).

Figure 7: Distance to shore map of the Netherlands (Dirksen 2015, p.20).

Although OK and UK show quite some resemblance, there are notable differences. Instead of making the sample variogram straight from the data, as was the case with OK, the UK interpolation method first removes a trend that can be fitted through the data of the logarithmic distance to the shore and the observed sample temperatures. Thereafter, the de-trended data is used to make the sample variogram in the same fashion as was the case with OK. Accordingly, the parameter selection and the variogram model fitting are also done identically.

Again, all the steps had to be conducted for every time step in order to make predictions and subsequently do a cross-validation, hence it was all automated with the Automap package in the R programming language (Hiemstra 2015).

Thin plate splines

As was stated in the theoretic background, spline interpolation methods fit multiple polynomials exactly through a set of data points. However, there are also 'cubic' splines that do not fit exactly through all data points but rather aim to capture the global variety (Hiemstra & Sluiter 2011). Figure 8 illustrates this concept, with on the left a spline that captures the global variety and on the right a spline that captures local variety.

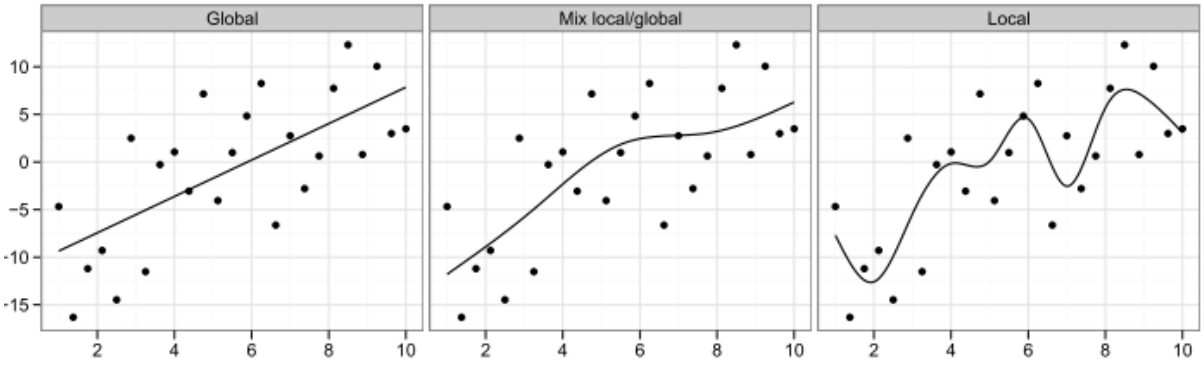


Figure 8: Splines fitted to random data focusing on either global or local accuracy (Hiemstra & Sluiter 2011, p.12).

Subsequently, Thin Plate Splines (TPS) are two-dimensional cubic splines that balance between local accuracy and global accuracy (illustrated by the middle spline in figure 8). Previous research has shown that TPS can be used to successfully interpolate temperature in the Netherlands, hence it is also tested in this thesis (Dirksen 2015; Sluiter 2012; Sluiter 2009). Furthermore, TPS can be compared to fitting a thin metal sheet through a set of data points. The balance between the local- and the global accuracy is established by using a cost function that on the one hand minimalizes the minimum error at the observation location, while on the other hand minimalizes the bending of the sheet. The degree to which this cost function favours the first over the second factor is controlled by the λ parameter. Accordingly, the λ parameter is estimated from the observation points by doing a Generalized Cross Validation (GCV). This procedure was also automated with R, with a script written by Hiemstra (Appendix E) that primarily uses the 'Tps' function from the 'Fields' package (Douglas et al., 2016; Hiemstra & Sluiter 2011). As a result, this method could be repeated for every time step in order to make interpolations and subsequently do cross-validations for the whole study period.

When?

Next to the selection criteria and the chosen interpolation methods, it also necessary to discuss the temporal resolution of this research step. Firstly, it should be noted that the WOW-NL dataset consists of temperature measurements that are registered every ten minutes. Next to the WOW-NL dataset, the official AWS measurements are also available for this resolution. It should be noted that this is considered as a relatively high temporal resolution in terms of meteorological standards.

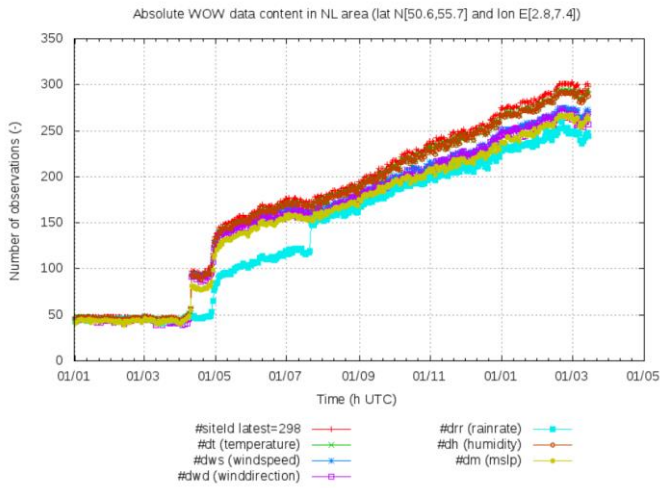


Figure 9: The growth of registered WOW-NL stations over time.

Besides that, the amount of amateur weather stations that are registered to the WOW-NL application is continuously growing (figure 9). However, it should be stressed the WOW-NL application only exists since April last year. Consequently, this means that the WOW-NL data is only available for 3 out of 4 seasons, which is an important limitation. Nevertheless, in order to capture the influence of the three different seasons in the interpolation results, it is chosen to conduct interpolations for the months: August, October, and January.

Coming back to the tested interpolation methods, this means that for every month temperature interpolations and cross-validations had to be made for approximately 4.000 time steps. Given the amount of cross-validations that are conducted, it is possible to plot the performance of the interpolation through time in order to get a comprehensive overview of the stability of each interpolation method.

In order to start with the next research step, the procedure of finding the best interpolation method had to be completed first. Accordingly, it was concluded that TPS was the most optimal interpolation method for this dataset. In order to read how this was concluded the reader is referred to paragraph 5.1.1.

4.1.2. Comparing interpolations with WOW-NL data

When the most optimal interpolation method for the selected dataset has been determined, the second part of the data assessment can begin. Accordingly, this part starts with conducting the TPS interpolation method for every time step. This is done in order to get estimated temperature values, based on the AWS measurements, for all the WOW-NL stations (locations). As a result, every WOW-NL measurement includes an interpolated temperature value as well as an observed temperature value. These values play an important role, as the main objective of this part of the research comprises the comparison of these two values.

Comparison methods

Firstly, it should be noted that a few different causes could result in differences (residuals) between the WOW-NL measurements and the interpolations. On the one hand, there are the combined factors that altogether form the bias of amateur weather station, and on the other hand there is the actual variation in temperature that exists across space. It should be stressed that is impossible to distinguish

these two with absolute certainty. However, the size of the residuals and their occurrences among different subsets of the data, gives insight into which factors cause issues for amateur weather stations. As a result, the used comparison method is dedicated to extract patterns from the bulk of (all) data by analysing different sub-sets (e.g., day versus night residuals, summer versus winter residuals). By using these different subsets it becomes clear which dimensions of the WOW-NL data are characterised by the biggest deviations from the AWS predicted temperatures.

Observed vs. predicted

Since there are approximately 3.4 million measurements compared to their associated interpolated temperatures, there are many useful ways of analysing and highlighting these differences. The first method that is used consists of plotting the measured temperatures against the predicted temperatures. This results in a comprehensive graphical overview that clearly illustrates the distribution and size of the residuals. Since this is done before corrections took place, they also show where outliers occur. Besides that, it should be noted that plotting the observed values versus the predicted values could be done for each and every subset.

Boxplots

Since plotting the observed values versus the predicted values merely gives a visual overview of general patterns, it is necessary to analyse the residuals further. Accordingly, this will be done according to boxplots of the residuals, which come with a few further benefits. Firstly, they provide some information regarding the symmetry and the skewness of the data. Secondly, boxplots clearly illustrate the spread and the outliers of the residuals. Furthermore, boxplots can be used to graphically display the differences between subsets since they can be plotted next to each other, which enhances the inter-comparisons between them.

Raster diagram

In order to gain a comprehensive overview of how the average and the mean residual fluctuates over time, it is chosen to plot raster diagrams. These include the mean residual for every hour and every month for all WOW-NL data. Accordingly, it should be noticed that these raster diagrams comprise three dimensions of data. On the x-axis the months are plotted, while the y-axis includes the different hours of the day. Furthermore, the cells of each raster are filled according to their associated mean residual values according to a colour ramp. Concluding, this plot will compactly summarize the deviations of the WOW-NL measurements according to different moments in time.

Bubble plots

Besides the statistical approach, it is also important to examine the spatial component of the data. As a result, bubble plots were made that illustrate the average and the absolute average residual per station on a map. These bubble plots illustrate whether the residuals per station are distributed across space relatively random, or if they are concentrated in certain places. Besides that, these bubble plots give a clear overview of which stations are close to the interpolations and which are not. Additionally, it is also becomes clear if the individual stations generally measure lower or higher than the interpolations.

Subsets comparison

The different plots of the residuals should be considered as the means to analyse differences between the measured temperatures by WOW-NL stations, and the interpolations of AWS measurements. However, instead of analysing all the measurements together, it is of considerable importance to analyse different subsets and compare the results. Accordingly, the following subsets were used.

Different seasons

The first subset consists of the six different months (July, August, October, November, December, and January) during three different seasons. The primary reason for the analysis of this subset is due to the fact that the different seasons are characterised by a considerable difference in temperature and radiation. According to Bell (2014), the different types of amateur weather stations can show different biases under different climatologic circumstances. Consequently, making different subsets for the different seasons illuminates to what extent the residuals fluctuate during seasonal change.

Day and night

One of the most important findings that Bell (2014) described includes the influence of solar radiation on the bias of amateur weather station observations. As a result, it is imperative to analyse weather the difference between the observed temperature and the predicted temperature is bigger at night than during the day. This is important since at night there is no solar radiation that directly hits the amateur weather stations, which means that this radiation bias should be nullified. Besides that, the radiation bias should also be bigger in the summer, as the sunlight is much stronger than in the winter. Therefore, the seasonal change has also been integrated in this subset. As a result, it becomes clear to what extent this radiation bias applies to the WOW-NL data in the selected months.

WOW-NL Station rating

Another interesting subset that has been analysed consists of the different WOW-NL station ratings (Appendix C). The station ratings refer to a quality classification system that the WOW-NL application has implemented, that gives amateur weather stations a rating between 1-5 (worst to best), according to their location attributes. The methods that are used to rate the location attributes are based on the standards that are prescribed by the COL (Climatological Observers Link), the WMO (World Meteorological Organisation), and the Met Office. The way the classification system works is shown in Appendix C. Furthermore, this subset is examined in order to test if WOW-NL stations that adhere to meteorological standards observe temperatures that are closer to the interpolations.

WOW-NL Urban Climate Zone rating

Next to the overall station rating, it is also relevant to consider a few individual factors from the WOW-NL classification system since they are of considerable importance to temperature measurements. The Urban Climate Zone (UCZ) index, classifies the stations surroundings according to the amount and type of built environment in its vicinity (Appendix C). This is especially relevant since urban climates tend to be warmer than rural climates (Santamouris 2014). The expected bias would thus be relatively higher for those stations located in urban areas since the interpolation model does not consider the urban environment in its predictions.

WOW-NL Exposure rating

The Exposure rating takes obstructions that are in the vicinity of the weather stations in to account. This is also an element that is important to analyse separately, since temperature measurements can be significantly influenced by surrounding buildings and constructions. Accordingly, these can shelter the weather station from wind, solar radiation, and rain which have a significant influence on temperature measurements. Accordingly, the stations which have the best exposure rating are expected to resemblance the interpolated temperatures the closest, as automatic weather stations also have maximum exposure.

WOW-NL temperature rating

Furthermore, the WOW-NL classification system also includes a specific rating for temperature instruments, which has also been used as a specific subset. This classification system is primarily based

on the calibration history and radiation screen. As was discussed, radiation has a large impact on temperature measurements. Consequently, the stations with high ratings in this regard are expected to have measurements that resemble the interpolated temperatures the closest.

Individual Performance

The next subset that is important to analyse comprises the individual weather stations. Since to date there are approximately 250 amateur weather stations subscribed to the WOW-NL application, it is possible to analyse the residuals per station. This is especially beneficial for making correction as stations with abnormal observations stand out by plots with predominantly gross errors and outliers. Furthermore, this also gives insight into whether individual station can perform well, and to what extent individual stations influence the total residuals of all WOW-NL stations.

Reputable WOW-NL stations

According to a field study where the most popular amateur weather stations were tested next to an official weather station from the Met Office, two type of stations from the same brand came closest to the official measurements (Bell et al., 2013; Bell 2014). These two stations include the Davis Vantage Pro 2 and the Davis Vantage Vue. As a result, it is interesting to test whether these findings are also valid for the interpolations and the measurements of WOW-NL stations. Accordingly, this subset includes the WOW-NL stations that are presumably either one of these two stations. Since it is not mandatory to disclose the station type when connecting a weather station to the WOW-NL application, it is necessary to filter these stations out otherwise. It is therefore chosen to retrieve model type indications from the columns '*Site Description'* and '*Additional information'* in the WOW-NL dataset. Stations belonging to this subset include one of the following words in either of these columns: Davis, Vue, VP2, VP, and Vantage.

Together, the different comparison methods and the different subsets give an overview of how close the WOW-NL data comes to reliable temperature estimates, based on official measurements. Furthermore, the comparisons also illuminate under what circumstances the WOW-NL stations experience difficulties in making plausible measurements. Additionally, these circumstances also give clear indications of what might cause these difficulties for the amateur weather stations. Altogether, this part of the research completes the methodological description of the data assessment. The results that come from conducting these steps are described in paragraph 5.1.2.

4.2. Gross error and outlier removal

When the assessment of WOW-NL data is completed, the data should be analysed for outliers and gross errors. This part of the research is crucial since the data assessment showed that the WOW-NL data includes quite some measurements that are highly unlikely and most probably the result of a malfunctioning WOW-NL station (paragraph 5.1.2). Besides that, other issues such as data gaps, repetitive stations, and stations that only include a few measurements also occurred in the WOW-NL data. If these measurements are not filtered out, they can influence the data integration negatively and cause unnecessary problems.

According to Osborne & Overton (2004), an outlier is defined as: "[..] a data point that is far outside the norm for a variable or population (Osborne & Overton 2004)". This means that the data point is so different from the rest of the population, that it might be generated by another procedure than a valid observation or data entry. Accordingly, it is important to understand that this type of measurements can have a significant influence on further analysis. For instance, using a wrong

temperature measurement in an IDW interpolation could result in a completely unrealistic map with false information. Hence, it is imperative to filter such measurements out.

However, according to Burke (2001), the removal of outliers is no trivial procedure. They argue that while a data point might be significantly different from the rest of the population, it can still be a valid observation. Nevertheless, there are some standard practices that can be used in order to determine which data points are outliers that can potential be removed. One of the most common methods to detect outliers consist of evaluating the mean of a variable minus/plus three times its standard deviation (Leys et al., 2013). Although, it should be noted that this method should only be used if the data follows a normal distribution. If the variable is completely normally distributed, this outlier detection method selects 0.13% of all the data and classifies them as outliers.

Although Leys et al., (2013) clearly underpin some limitations of this method, it is chosen as a starting point in this research (their objections merely apply to smaller datasets). Furthermore, it should be stressed that this method can only be used to identify possible outliers. In order to remove them there should be a valid reason, which is highly dependable on the type of data. For instance, temperature measurements above 60 ℃ in the Netherlands are not possible, so it is safe to conclude that this is an error. As a result, it is therefore necessary to make rules for selecting and removing outliers and gross errors for the WOW-NL dataset.

4.2.1. WOW-NL outliers and errors

Firstly, it should be noted that in order to filter out the outliers and gross errors two variables are evaluated. On the one hand, the residuals per measurement are taken into account, while on the other hand the observed temperature values are considered. Since the results in paragraph 5.1.2 illuminated a few issues which are unique for this dataset, the following objectives are set:

- Filtering out repetitive measurements
- Filtering out stations that were off most of the time
- Filtering out unrealistic measurements

Filtering out repetitive measurements

The first objective is dedicated to WOW-NL stations that are repeating measurements for a certain amount of time. The data assessment clearly showed that some stations (e.g., 936386001, 936496001, and 938336001) have repetitive measurements while the actual temperature was most likely varying (Appendix D). Accordingly, it should be assumed that these measurements are incorrect, since in reality temperatures do not stay stable for long periods of time. However, it should be stressed that it is not aimed to remove all measurements from a station if it has been in a repetitive state for a given period of time, since other measurements can still be correct. In this research it is therefore chosen to remove all measurements that are repeated for at least 3 hours (18 time steps). This means that for every station the consecutive measurements have been analysed.

Filtering out stations that were off most of the time

The second objective is dedicated to the stations that only show a few measurements in the whole study period. As can be observed in Appendix D, some stations (e.g., 927226001, 933846001, and 923406001) have only a few or no bins with observed versus predicted temperatures. Regardless of their accuracy, it is chosen to assume that these stations do not provide reliable measurements due to their poor sample size. Consequently, all WOW-NL stations that provide measurements for less than 5% of the total study period are removed from the dataset. With this threshold it is taken into account that some station have only joined WOW-NL in January. These will not be removed as 5% of

the time accounts for approximately 9 days. Furthermore, the data gaps will be removed by keeping all the measurements that do not include an observed temperature out of the dataset.

Filtering out unrealistic measurements

The unrealistic measurements are filtered out by evaluating the temperature residuals. As was mentioned, the standard practice of using the mean minus/plus three times the standard deviation was the starting point. Although, before this method could be applied, it was necessary to examine the distribution of the residuals since the method is only applicable to normally distributed variables. The next step consists of determining whether or not the outliers seem realistic. In this regard it is important to keep in mind that the residuals include the difference between temperature estimations based on automatic weather stations and the observations that are made by WOW-NL stations. Since microclimates can deviate significantly from the air temperature on a coarse scale at which the KNMI measures, it is important to have relatively large boundaries for the bias. Consequently, it is chosen to not remove data with a bias of -10℃ to 10℃, and all data that has a residual smaller than the mean minus/plus three times the standard deviation.

After all the previously described steps had been conducted, the dataset was ready for the next research step, which includes the data integration. This was done according to three different data integration scenarios, that each aim to explore and illuminate potential benefits that arise with the integration of formal and informal data. Finally, the results of the gross error and outlier removal can be observed in paragraph 5.2.

4.3. Data integration scenarios

The final part of this research consists of the data integration. Accordingly, the informal data and the formal data are integrated in order to explore its potential benefits. It was chosen to use three different scenarios which each aim to illuminate potential benefits, by integrating the data in a different way. The three scenarios include: (1) threating the WOW-NL data as equal compared to the AWS data, (2) using the WOW-NL data only as a secondary predictive variable, and (3) making further corrections for solar radiation to the WOW-NL data. The methodological steps that are carried out during these scenarios are described in the following paragraphs.

4.3.1. WOW-NL as equal compared to AWS

The first scenario consists of treating the WOW-NL data as equal to the AWS data. To formulate this scenario more straightforward, it consist of ‘doing nothing’ to the WOW-NL data and treating it as if it are accurate measurements. As a result, the WOW-NL data can immediately be integrated and used for further analysis. However, the success of integration still needs to be tested.

This is done according to two different analyses. Firstly, the success of the integration is tested according to interpolations and quality indicators. Accordingly, the same methods (including parameters) that were used in paragraph 4.1.1 are now tested with the integrated data. This means that both WOW-NL stations and automatic weather stations will be considered as equal, and both be taken into account in the interpolations. However, the study period is slightly shorter than in paragraph 4.1.1, due to the intensity of the calculations. It is therefore chosen to make interpolations and cross-validations for all time steps for one week in August, October, and January. This way the different seasons are captured in this analysis as well.

However, instead of doing the leave-one-out cross-validation for all stations, it is only done for the automatic weather stations. This means that in the cross-validation, the values of the automatic

weather stations are predicted according to all other stations (both WOW-NL and AWS). This is due to the fact that these are the only measurements that are validated, hence it is the only comparison that makes sense. As a result, this scenario examines if the AWS measurements can be predicted more accurately according to integrated data, or according to AWS data only. It is predicted that WOW-NL stations might improve predictions especially when there is much spatial variation which is hard to capture with fewer stations. Again, the RMSE is plotted over time, and compared to the RMSE of the interpolations that were done in paragraph 4.1.1.

Besides the evaluation in terms quality indicators, it is also necessary to examine the result of the interpolations visually. Since the observation sample has improved drastically by the data integration, the resultant maps are affected as well. Therefore, the different interpolation results are also compared in terms of their visualisation. However, this is only done on a small scale (one example each) since it is not feasible to compare thousands of maps. Finally, the results of this scenario can be observed in paragraph 5.3.1.

4.3.2. WOW-NL as secondary predictive variable

The second integration scenario threats the WOW-NL data as less reliable than the AWS measurements. This is done by using the WOW-NL data only as a secondary predictive variable in the UK method as was described in paragraph 4.1.1. This means that the WOW-NL data will be used to remove a trend that can be fitted through the WOW-NL data and the observed temperatures at automatic weather stations. Thereafter, the de-trended data will be used to make the sample variogram as was the case with OK. However, this means that a continuous surface with temperature values based on WOW-NL measurements is necessary in order to have a temperature estimates on the location of automatic weather stations. As a result, this scenario consists of (1) making interpolations based on the WOW-NL data, and (2) conducting interpolations and cross validations with the WOW-NL data as secondary variable in the UK method.

WOW-NL interpolations

In order to use the WOW-NL data as a second variable in the UK interpolation method, it is required to have temperature estimation based on the WOW-NL data for the automatic weather stations. In order to obtain those values it is chosen to use the TPS interpolation method. This is chosen due to the overall robustness of the TPS method, combined with the limited time available for this research. Ideally, another study like was done in paragraph 4.1.1 would be necessary to find out which interpolation method performs the best with more observation samples. However, since the TPS method has proven itself as a well performing, robust method for interpolating temperature data with a high temporal resolution, it is justified to use this method. As a result of this step, all AWS measurements have both the temperature that was observed by the official station, as well as a temperature value based on the WOW-NL interpolation. This makes it possible to make interpolations with the UK method, and the WOW-NL data as secondary variable.

UK cross validation

As was the case in the previous scenario, it is necessary to conduct interpolations and subsequently test the quality. Accordingly, it is chosen to also perform the leave-one-out cross-validation. Since the WOW-NL data is merely used as a secondary predictive variable, the cross-validation can only be done for the automatic weather stations. This means that again the RMSE can be plotted against time, so that a comprehensive overview of the interpolation quality can be obtained. Besides that, it should be stressed that the study period is exactly the same as in the previous scenario. As a result, the outcome of the first scenario can also be compared with the results of the first scenario. Furthermore, the results of the second scenario are also interpreted visually, with maps that illustrate the result of the interpolations. It is chosen to use a time step that illustrates a situation where the interpolation outperforms the AWS TPS interpolation.

4.3.3. Correcting WOW-NL observations for solar radiation

The third and final integration scenario aims to correct the WOW-NL data before it is actually integrated. Since Bell (2014) states that incoming solar radiation has a significant influence on the bias of the WOW-UK measurements, it is relevant to examine if this also applies to the WOW-NL data. This is examined by a prediction model that aims to predict the interpolated temperatures based on the WOW-NL observed temperature values and the incoming solar radiation. Accordingly, this made it possible to test whether the solar radiation has any predictive significance for the residuals between the interpolated temperature values and the observed WOW-NL temperatures. This research step is done according to two different prediction models. Firstly, a multiple regression analysis was conducted for all the data, and secondly, a random forest model was used to analyse specific days where there was relatively much solar radiation. However, before the prediction models could be constructed, reliable estimations of solar radiation at the sites of WOW-NL stations had to be made. As a result, this integration scenario consist of three steps: (1) interpolating radiation data, (2) a multiple regression analysis, and (3) a random forest analysis.

Sun interpolations

Since the WOW-NL station do not measure radiation, there is a need for another source that includes measurements for solar radiation at any location in the Netherlands. Ideally, this should be derived from satellite images (e.g., cloud cover) or weather forecasting models. However, due to the limited time available for this research it was chosen to make interpolations from hourly averages at automatic weather stations. Since the solar radiation is merely used as a global predictive variable, this rough estimates were viewed as sufficient. Consequently, the AWS measurements of average joule per square centimetre were used to make interpolations (KNMI 2016a). This means that the temporal resolution for this variable is slightly lower than the WOW-NL resolution. Since rough estimates were valued as sufficient, it was chosen to use the IDW interpolation method with the standard power function (q=2). The selection for IDW was mostly made due to the robustness of the IDW method. As a result, all WOW-NL measurements also include an estimated joule per square centimetre which could be used as input for the prediction models.

Multiple regression analysis for all data

In order to describe the relation between an ‘dependent’ variable and a related ‘predictor’ variable, a linear regression analysis can be used (Montgomery et al., 2015). In case there is more than one predictor variable, the highly similar multiple regression analysis is a suitable alternative. Accordingly, the multiple regression analysis aims to fit a linear model through the data, which best predicts the dependent variable according to the predictor variables. The dependent variable is estimated by: (1) a constant, (2) and a slope multiplied with the predictor variable(s). In order to make a prediction

model that generates good estimates of the interpolated temperatures, the variables: observed temperatures, incoming solar radiation, and the time of measurement were used. This way it can be tested whether sunlight is a good predictor variable for the deviation from the observed temperature by WOW-NL stations. This resulted in the following formula for the multiple regression analysis:

Predicted temperature = b0 + b1(solar radiation) + b2(observed WOW-NL temperature) + b3(moment of measurement).

The resultant model can be used to correct the WOW-NL observations. However, it should be stressed that the predicted temperatures play an important role in this model. All the variance that can be explained according to the model will be removed and the residuals will be the only temperature deviation from the interpolations that are left. In theory, the resultant deviation from the predicted temperatures should be viewed as the local variation in temperature. Nevertheless, it would not make sense to do cross-validations in order to test the quality of the corrected data, since the interpolations were already based on the AWS data.

Random forest model for hot days

Besides the multiple regression analysis for the whole dataset, it is also interesting to use a more powerful prediction method for a few exemplary sunny days. Accordingly, it is chosen to use the random forest regression method for two relatively hot days in August. The random forest method is an ensemble machine learning approach, which means that it uses many (weak) models in order to construct a (powerful) model by aggregation (Breiman 1999; Liaw & Wiener 2002). The random forest model is an ensemble of decision trees, which generates either a classification or a predicted value (regression). The threes are individually constructed by randomly selecting a subset of the training data, and subsequently the nodes are split by using a random selection of predictor variables. Ultimately, the different threes together form the forest, and the prediction value is chosen by the average outcome of all the threes combined. The random forest model is also selected since it is suitable for dealing with larger datasets (Breiman 1999; Liaw & Wiener 2002). Finally, it should be noted that both correction models are supported with maps that are based on the corrected WOW-NL data.

5. Results and discussion

This chapter includes all the results that came from the different steps that were conducted during the research. These include: the data assessment, the gross error and outlier removal, and the data integration. In order to keep the structure of this chapter clear, the results of these different steps are described separately.

5.1. Data assessment

The data assessment part consists of two steps. Accordingly, the first step consists of finding the most optimal interpolation method for AWS temperature measurements, in order to gain accurate temperature estimates at the sites of WOW-NL weather stations. Next, the second step includes an extensive comparison between interpolated temperature values, and the actual temperature values that were measured by the WOW-NL stations. These two research steps together form the methodological part of the data assessment and are described in the next two paragraphs.

5.1.1. Finding the best interpolation method

In this part of the research, four interpolation methods were compared for an extensive period of time. More specifically, the interpolation methods IDW, OK, UK, and TPS were tested for three different months in different season (August, October, and January). This period was chosen in order to find out how the interpolation methods perform under different climatic circumstances. Furthermore, the temporal resolution is also relatively high, as the interpolations and cross-validations were made for every 10 minutes during the three month period. Subsequently, this made it possible to plot the RMSE of every cross-validation against its moment in time (figure 10).

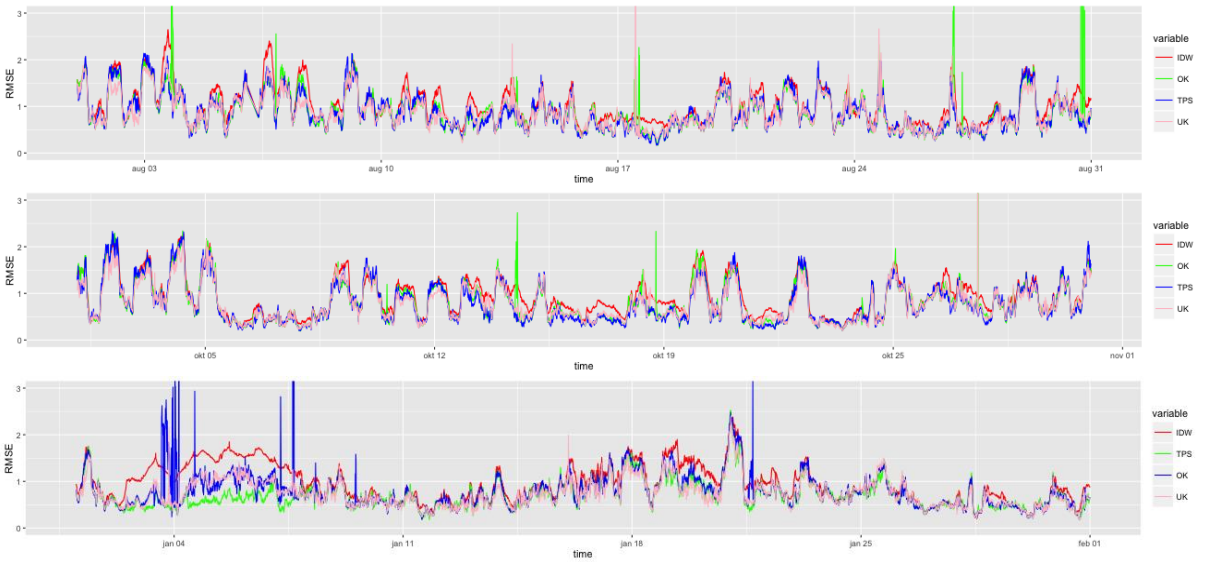


Figure 10: The RMSE of cross-validation from the four tested interpolation methods IDW, OK, UK & TPS based on AWS in August, October & January (y-axis is fixed at RMSE = 0-3.0).

The results of the cross-validation clearly show big differences between the four interpolation methods. Firstly, it should be noted that IDW structurally performs less than the other three methods in terms of RMSE values; this is valid for all three months. Furthermore, it becomes clear that the two Kriging methods tend to be less stable, as they have relatively high, and much outliers compared to the other two methods. It should be noted that the y-axis figure 10 is fixed to RMSE values 0-3,

however the outliers of the Kriging methods occasionally go up to an RMSE of approximately 15 (Appendix A). Evidently, these extreme outlier interpolations are useless when compared to the WOW-NL measurements since they do not produce reliable temperature estimates at the site of WOW-NL stations. Furthermore, it should be noted that TPS shows good performance on both stability and RMSE values.

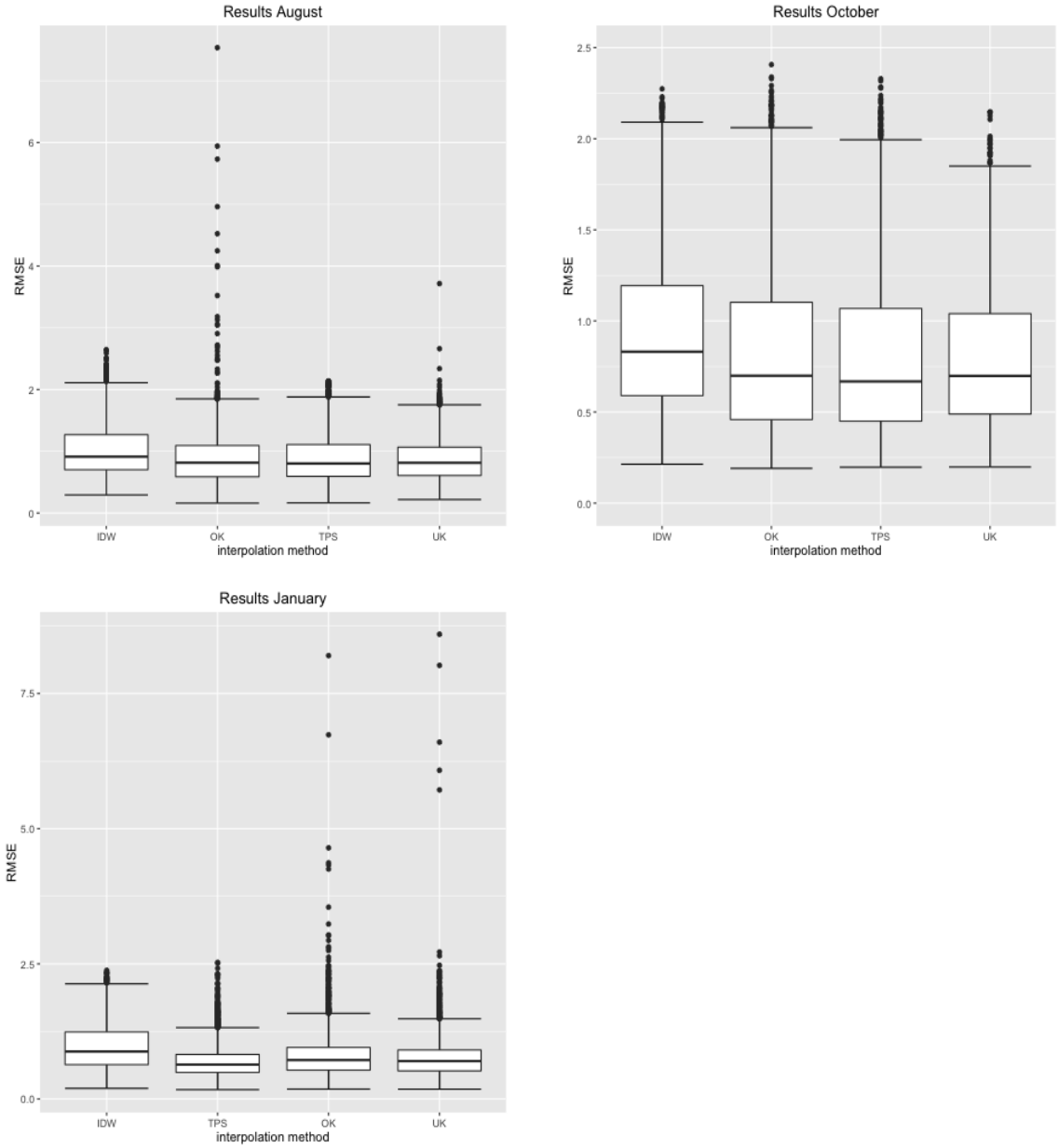


Figure 11: RMSE Boxplots of cross-validation from the four tested interpolation methods IDW, OK, UK & TPS in August, October & January.

Besides that, the instability of the Kriging methods is also confirmed when examining the boxplots that come from the cross-validations (figure 11). The y-axis for the boxplot for October is fixed to 0-2.5 since a couple of outliers on 28-10-2015 have an RMSE of approximately 15, which make the figure unreadable if they are included (Appendix B). Nevertheless, the boxplot from January clearly shows that the Kriging interpolations have more, and bigger outliers than IDW and TPS.

When exploring the Kriging outliers further, it should be noted that it is known that a small number of sample observations can lead to difficulties when making the variogram. According to Hengl (2007),

the 35 automatic weather stations are not enough to make reliable variograms, as he advises a minimum of 50 sample observations. Comparing his advice to the results that are presented in figure 10 and 11, it should be acknowledged that the Kriging interpolation methods indeed seem to perform poorly in some situations. However, the vast majority of the interpolations are successful. So, if only one, or only a few interpolations had to be made, the Kriging methods would still be worth testing.

Besides that, it is interesting to further evaluate Kriging interpolations with poor quality. Accordingly, figure 12 illustrates interpolated temperature maps made with TPS and OK on a moment that the Kriging interpolations perform poorly. It is clear that the general patterns look the same, however the OK interpolation has some strange deviations in the southern part of Limburg. There is no reason to assume that temperatures deviate approximately 30 degrees Celsius within such a small area, hence these estimates are most likely false. When these two interpolations were examined by means of a cross-validation, TPS significantly outperformed OK with an RMSE of 0.6 as opposed to an RMSE of 7.5. Concluding, these maps illustrate that comparisons between the interpolated values and WOW-NL measurements do not make sense if the interpolations have poor quality.

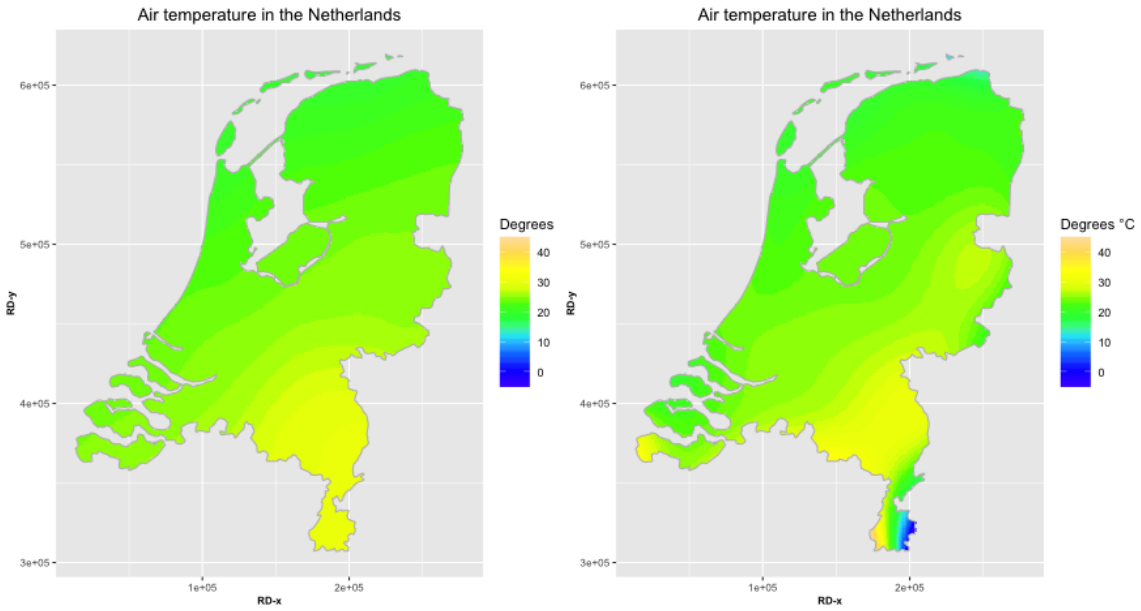


Figure 12: Air temperature on 30-8-2015 17:20 interpolated with different methods, from left to right: TPS & OK.

Coming back to the overall comparison between the four interpolation methods, it is also important to examine the performance in terms of average RMSE and ME. Table 2 clearly illustrated that the TPS interpolation method performs best overall, however in August and October it is similar or less compared to UK. Nevertheless, it should also be stressed that the differences are relatively small. Accordingly, the difference between the best and the worst interpolation method is only 0.17 degrees Celsius in terms of RMSE. Besides the RMSE, it is also important to check if interpolation methods structurally over, or under predict. Table 3 shows that this is not the case, as all interpolation methods have a ME of approximately 0%.

	August	October	January	Total
IDW	1.01	0.92	0.95	0.96
OK	0.89	0.82	0.79	0.84
Uk	0.87	0.80	0.74	0.80
TPS	0.88	0.80	0.70	0.79

Table 2: Average RMSE values

	August	October	January	Total
IDW	0.02	-0.01	-0.04	-0.01
OK	0.01	0.00	0.00	0.00
Uk	0.01	0.00	-0.01	0.00
TPS	0.02	0.01	0.00	0.01

Table 3: Average ME values

Concluding, the most appropriate interpolation method is chosen according to the selection criteria (stable, replicable, high quality) that were described in paragraph 4.1.1. In terms of stability the interpolation methods IDW and TPS considerably outperformed OK and UK, as they had way less, and lower outliers.

Furthermore, the complexity of the OK and UK methods is also clearly higher since they require variograms, including many parameters that have to be selected. Besides the Kriging methods, IDW is viewed as the least complex method since it only considers distance between measurements according to a simple power functions (q), and radius. The TPS method is somewhere in between because the polynomial functions and the cost function are more complex from a mathematical standpoint.

Finally, the quality of TPS in terms of RMSE was clearly better than all other methods. Besides TPS, UK and OK performed more or less equally, with UK performing slightly better than OK. The IDW method structurally performed the worst in this regard.

When considering all three selection criteria it should be stressed that TPS is the most appropriate interpolation method for this dataset. It is the best performing in terms of quality, while it is also a stable method that is not too complex. The only real alternative could be IDW since it is a less complex method that is also stable. However, it structurally performs less than TPS while its relative simplicity does not impose a clear benefit over TPS that outweighs it shortcomings in terms of quality. Besides that, the extreme outliers of the Kriging methods make them unusable for some time steps in the dataset. As a result, TPS is chosen to make interpolations to the WOW-NL stations in the next research step.

5.1.2. Comparing interpolations with WOW-NL data

The second part of the data assessment consist of the extensive comparison between interpolated temperature values and the observed temperature values by the WOW-NL stations. Accordingly, the first step consists of making interpolations for every time step for the whole dataset. As was decided in the previous paragraph, this is done with the TPS method. As a result, all measurements by the WOW-NL stations include both an observed temperature as well as a predicted temperature.

All WOW-NL data

An overall plot for all time steps and all WOW-NL stations gives a first indication of how the WOW-NL data (approximately 3.4 million observations) resembles the predicted temperatures (figure 13).

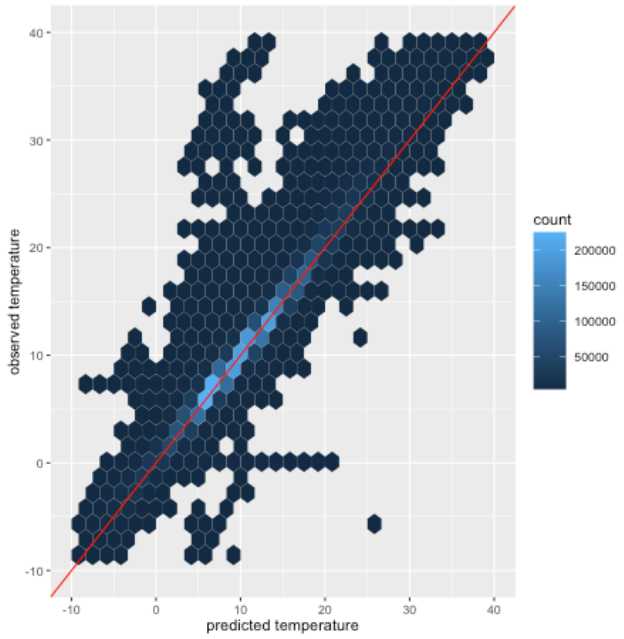


Figure 13: Hexagonal binned data points of predicted temperature (KNMI interpolations) vs observed temperature (WOW-NL) for all time steps in July, August, October, November, December, and January.

Accordingly, it becomes clear that in general most WOW-NL stations observe higher temperatures than predicted temperatures. However, it should be noted that there are also many cases in which they observe lower temperatures. Moreover, between 0°C and 20°C, there is a large group of WOW-NL observations that closely resemble the interpolated values. Besides that, it is hard to extract further conclusions from this plot since it includes all measurements, from all WOW-NL stations, for approximately 26.000 time steps. Nevertheless, the plot also includes clear indications that there are some gross errors and significant outliers present in the dataset. For instance, the hexagon bins with an observed temperature of 0 °C strongly suggest that either one or a few WOW-NL stations are registering repetitive measurements. This is implied since they form a straight line, which can be the results of a station that keeps registering 0 °C while the actual temperature is varying. Besides that, there are some other suspicious hexagon bins at other observed temperature values. For instance, the one that is located at observed temperature -6 °C and predicted temperature 25 °C. This difference between the interpolated temperature and the actual temperature is highly unlikely, which strongly suggests that there are also quite some measurements that do not make sense and should not be taken into account for the data integration. Concluding, it is imperative to further analyse subsets of the data to gain better insight into the WOW-NL bias.

Figure 14 confirms the conclusion that most WOW-NL stations rather observe higher temperatures than lower temperatures, compared to the interpolations. Furthermore, it also becomes clear that there are quite some differences among the WOW-NL stations in terms of average residuals. Besides that, figure 14 gives a comprehensive overview of which stations perform very poorly and are likely to contain big errors. For instance the WOW-NL stations which are located in Zeeuws-Vlaanderen and in Walcheren show average residuals that far exceed 5°C. It is highly unlikely that the actual temperatures are deviating from the interpolations with that amount. As a result, these bubble plots indicate that the combined factors that altogether form the bias of amateur weather station are significantly influencing its measurements in some cases.

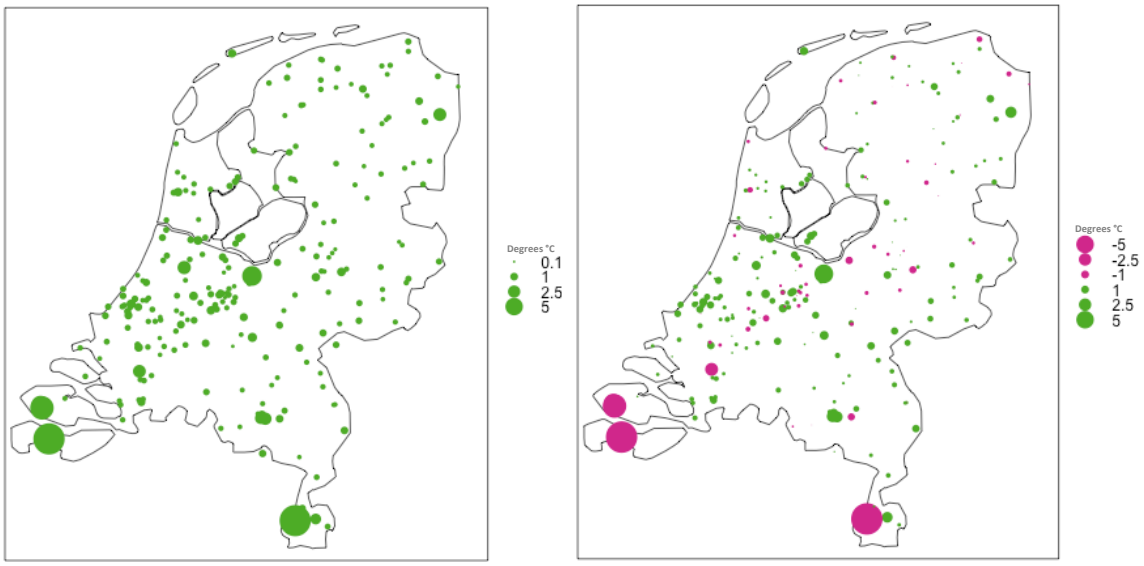


Figure 14: Average residual per WOW-NL station of predicted temperature (AWS interpolations) vs observed temperature (WOW-NL) for all time steps in July, August, October, November, December, and January. From left to right: Average absolute residuals, and Average residuals.

Individual stations

In order to examine what exactly causes these average residuals, it is necessary to analyse the residuals per individual station. Accordingly, these can be found in Appendix D, where for each individual WOW-NL station the predicted temperature is plotted against the observed temperature. The stations that caught attention in the bubble plots turn out to be stations that have measurements that are obviously wrong. Station id 936496001 (Walcheren) and 938386001 (Zuid-Limburg), are both stations that have repetitively been measuring 0°C for an extensive period of time. The WOW-NL station with site id 933846001 (Zeeuws- Vlaanderen) was only operable for a short period of time and also only registered 0°C (Appendix D). Further analysing these stations on the WOW website from the Met Office shows that these stations either do not provide measurements anymore, or are re-registered under a different name and or site id (Met Office 2016).

Beside the stations that caught the attention in the bubble plots, the individual plots also illuminate some additional trends. Firstly, it became clear that there are quite some stations that have unrealistic errors in them, as can be observed in the bottom plots of figure 15. As a result, it should be acknowledged that some form of gross-outlier and error removal is imperative for this dataset. Not doing so would clearly propagate these erroneous measurements towards the data integration, resulting in an unreliable integrated end product. Besides that, the middle plots illustrate that there are also WOW-NL stations that very closely resemble the interpolated values. Finally, the top plots in figure 15 show WOW-NL stations that increasingly deviate from the interpolated values as the predicted temperature starts rising. This strongly suggests that these WOW-NL stations are characterized by some degree of radiation bias as was described in paragraph 3.1.2. Furthermore, this would be in alignment with the findings of Bell (2014), as he describes that radiation bias is one of the most challenging obstacles for amateur weather stations. Nevertheless, this is only an indication, hence more decisive results are required.

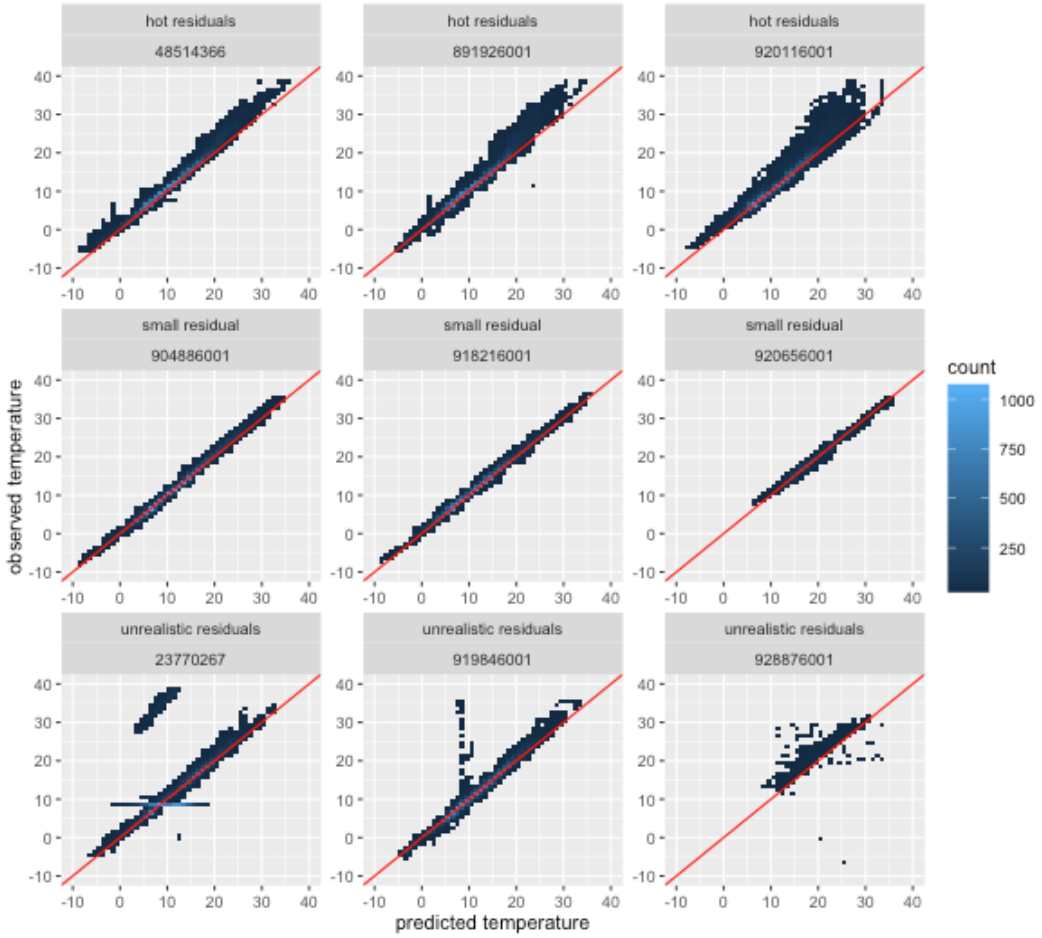


Figure 15: Individual WOW-NL stations with binned data points of predicted temperature (AWS interpolations) vs observed temperature (WOW-NL) for all time steps in July, August, October, November, December, and January.

Seasonal and temporal subsets

Besides the individual WOW-NL station, there are a few subsets that deserve some attention as well. Firstly, the seasonal differences are examined, as they include different temperatures, degrees of incoming radiation, and other climatological aspects that can influence WOW-NL measurements. Consequently, it is chosen to make subsets by season, which means that July and August comprise summer, October and November comprise autumn, and December and January comprise winter.

It becomes clear that the residuals of the WOW-NL measurements and the interpolations vary considerably across the different seasons (figure 16). In the summer, the residuals are the largest and show a sizable difference with both autumn and winter. Besides that, it should be noted that most of residuals are positive which means that the WOW-NL measurements are mostly warmer than the interpolated temperatures. Furthermore, the difference between autumn and winter is smaller than the gap with summer, although most of the residuals are still positive. Also, the range of the outliers is much bigger in the summer than in the autumn and the winter.

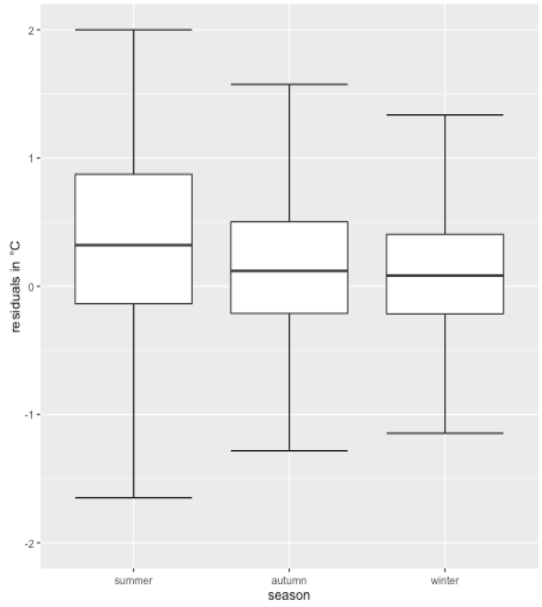


Figure 16: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to season (no outliers plotted).

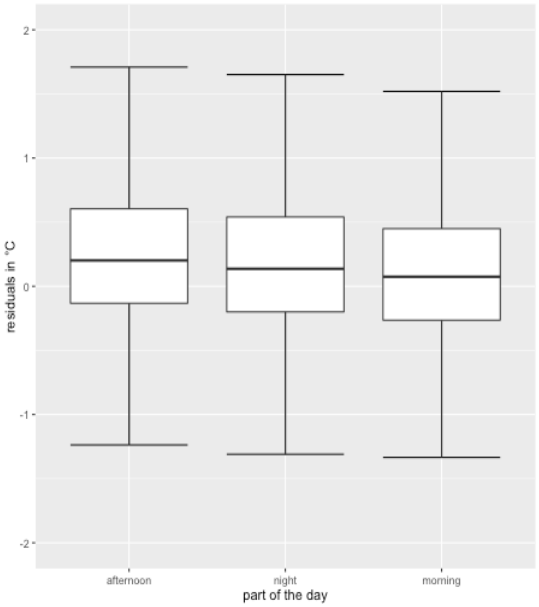


Figure 17 Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to the part of the day (no outliers plotted).

Another important subset that is analysed, consist of the difference between the parts of the day. It becomes clear that the residuals vary less compared to the seasonal change as the inter quartile range is much more similar (figure 17). However, there are some slight differences as the daytime shows the largest residuals, followed by the night, and subsequently by the morning. Nevertheless, it should be mentioned that the boxplot regarding the different times of the day is based on all the measurements in the different seasons; so seasonal differences could interfere with the comparison.

As a result, it is chosen to make raster diagrams that comprehensively show the variation of the residuals over time, and during the different seasons (figure 18). Again, this figure underpins the notion that there is a considerable difference in terms of residuals between the summer period and the winter period. Besides the conformation of the previous plots, these diagrams also show some new information. Firstly, it becomes clear that difference between average residuals in the afternoon and the morning are much higher in the summer than during the winter. This difference gradually evolves during autumn, as October still has clear differences between the afternoon and the morning, while November does not. Furthermore, the afternoon in the summer shows the largest average differences between the WOW-NL measurements and the interpolated temperatures. Besides that, the raster diagram also highlights a situation where the WOW-NL stations averagely measure lower temperatures than the interpolations, as the average bias in November is approximately -0.2°C. Finally, it should be noted that the differences between the WOW-NL measurements and the interpolation show considerable variation over time. It is however not clear what causes these fluctuations, as they are merely revealed and highlighted. Moreover, it is of critical importance to understand these aspects of the WOW-NL data so that they can be taken into account in the data integration.

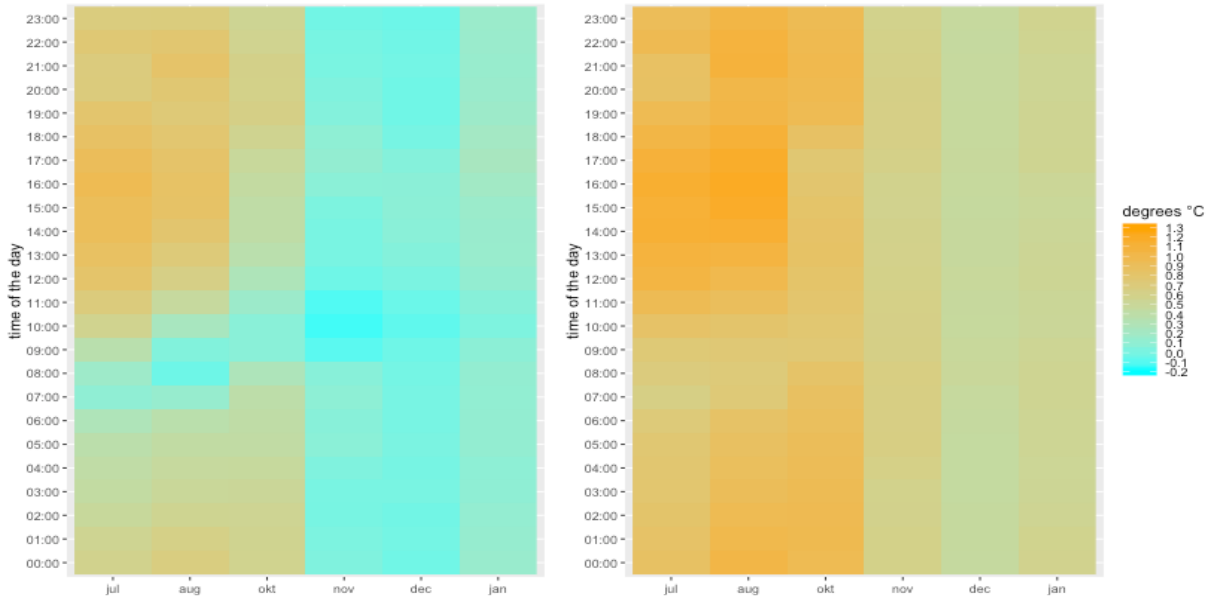


Figure 18: Raster diagrams of average (left) and absolute average (right) WOW-NL residuals of predicted versus observed temperatures, represented for every hour and every month (July, August, October, November, December, and January).

Station attribute subsets

Besides the temporal subsets, it is also important to examine the WOW-NL data according to their attributes (Appendix C). The most obvious subset consist of the difference in temperature instrument according to the WOW-NL standards. It would be logical to assume that the instruments with the highest classification (A) would show the best results, however this was not the case (figure 19). Conversely, the temperature subset did not seem to give any significant results as all categories had more or less the same average bias.

Furthermore, the subset that was made according to the exposure of the WOW-NL stations did come with some significant differences (figure 20). Accordingly, the first category (very open exposure), showed the closest resemblance with the interpolated values, although the differences with the next three categories is only small. Nevertheless, the category that includes the WOW-NL station that are very sheltered (category 5), shows the biggest deviation from the interpolated values.

The UCZ index comprises the next subset, and showed some variety in terms of average bias (figure 21). Firstly, it becomes clear that the rural areas and the areas with a small amount of buildings have the smallest average bias. Secondly, it should be noted that the differences between these categories (7-5) do not seem notable. The other categories do not seem to follow an obvious pattern in terms of average bias, as they vary substantially.

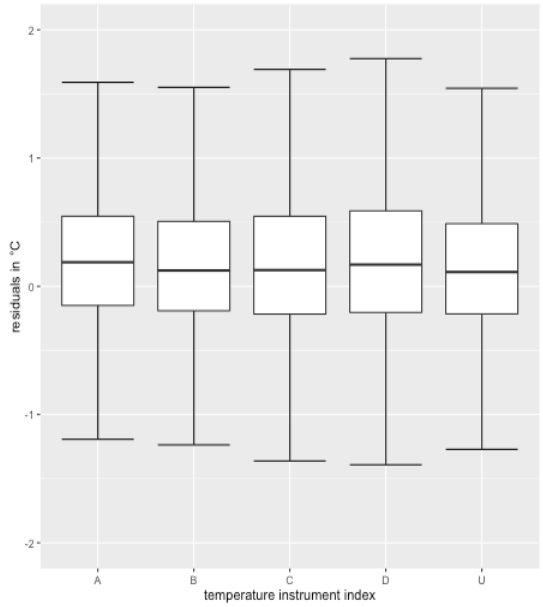


Figure 19: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to temperature instrument index (no outliers plotted).

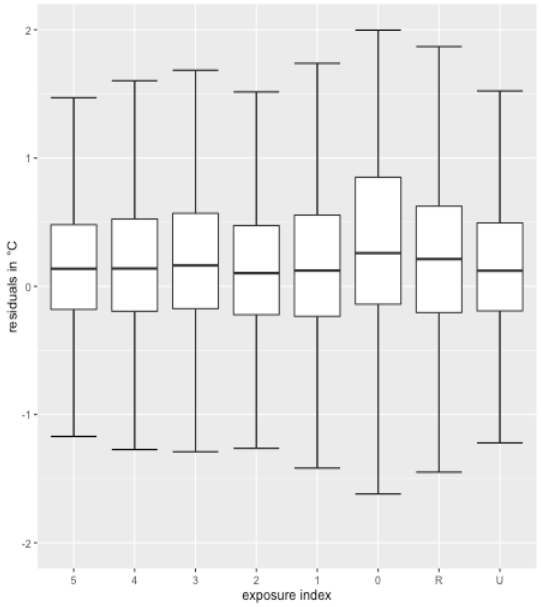


Figure 20: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to exposure index (no outliers plotted).

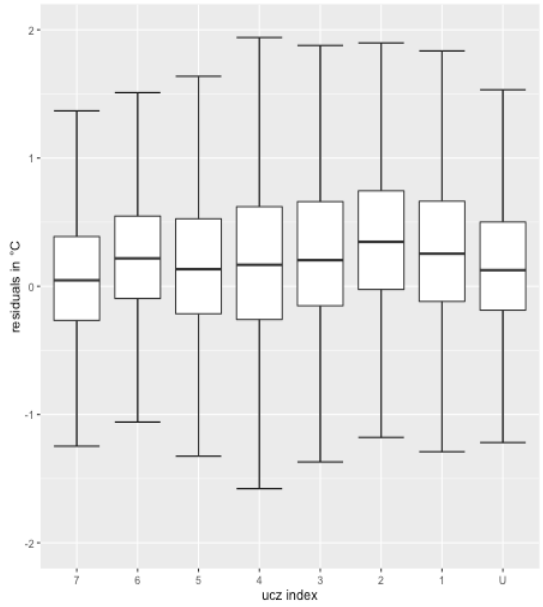


Figure 21: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to ucz index (no outliers plotted).

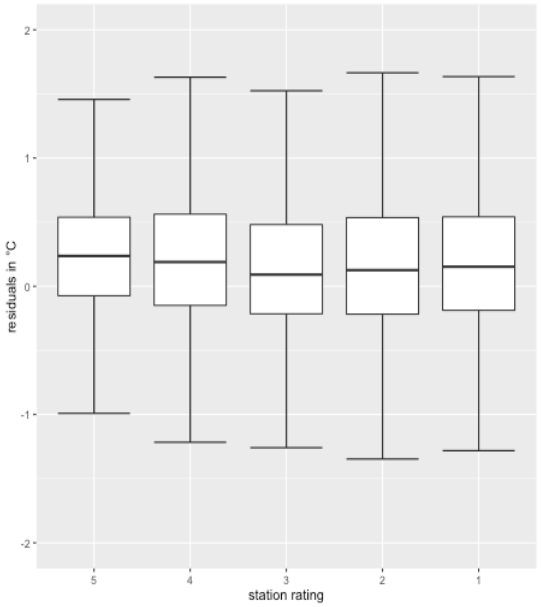


Figure 22: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to station rating (no outliers plotted).

The last attribute subset that was tested includes the overall WOW-NL station rating (figure 22). Although it was expected that there would be significant differences between the categories, they are rather similar in terms of average bias. This means that regardless of the WOW-NL station rating, the measurements deviate from the predicted temperatures in a similar fashion. Although, this similarity was not predicted, it is in alignment with the results from the previous subsets, as the station rating is partly based on them (Appendix C).

Reputable WOW-NL stations

Furthermore, the final subset that was analyzed was based on reputable weather station brands. Accordingly, the metadata was analyzed to filter out Davis weather stations. This was done since Bell (2014) found that the Davis stations resemble the measurements of official Met Office stations the closest as was described in paragraph 4.1.2. However, the results from the WOW-NL compression did not show similar patterns (figure 23). The Davis stations have a slightly lower mean deviation from the interpolations as well as slightly lower outliers. Nevertheless, these differences are small, and should be viewed as negligible.

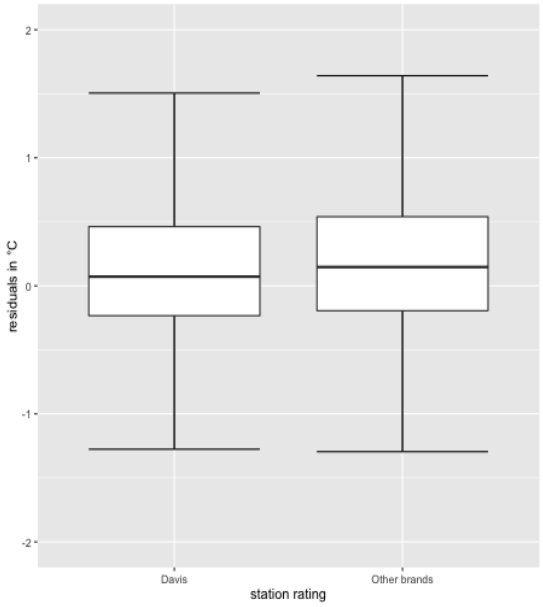


Figure 23: Boxplots of all WOW-NL residuals of predicted versus observed temperatures, grouped according to reputable brand (no outliers plotted).

5.2. Gross error and outlier removal

Now that the assessment of the WOW-NL data has been completed, it has become clear, which parts of de WOW-NL data deviate most from the predicted temperatures. Besides that, it has also been illuminated that there are quite some WOW-NL measurements that seem to be the result of a malfunctioning WOW-NL station or other error. As was described in paragraph 4.2, these erroneous measurements had to be removed in order to prevent propagation towards the data integration. The purpose of this research step is threefold, and consist of (1) filtering out repetitive measurements, (2) filtering out stations that were off most of the time, and (3) filtering out unrealistic measurements.

5.2.1. Filtering out repetitive measurements

The first step consists of filtering all the station out that were registering repetitive measurements. As was discussed in paragraph 4.2, it is chosen to filter out all observations that are similar for at least three hours (18 time steps). This was done by ordering all the data per individual station and sorting it by time step. Accordingly, the measurements that showed identical repetitive temperature measurements for 3 hours or longer, were removed from the data. The result of this step can be observed in figure 24, which shows the original data in the top left graph, and the results of the first step in the top right graph.

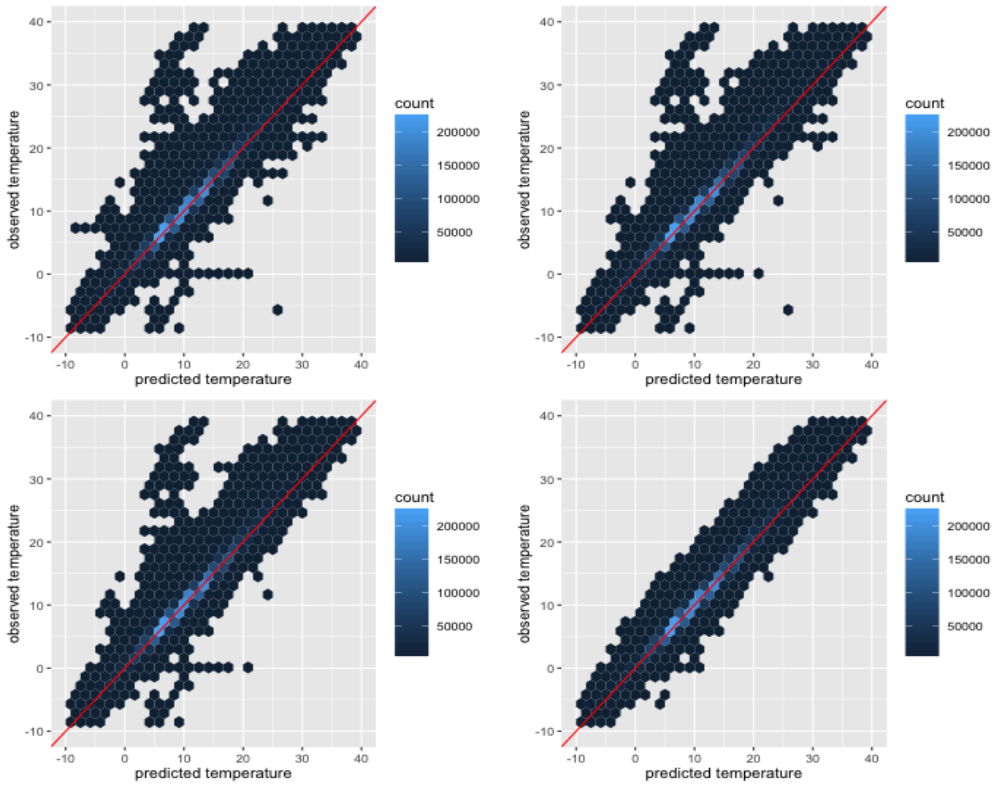


Figure 24: The predicted temperatures versus the observed WOW-NL temperatures: original data (top left), and after filtering out repetitive measurements (top right), filtering out stations that were off most of the time (bottom left), and filtering out unrealistic measurements (bottom right).

5.2.2. Filtering out stations that were off most of the time

The second step is dedicated to filter out all the measurements of stations that were on only a short amount of the study period. This was done in order to filter out all stations that only provide a small sample, which are not viewed as reliable. This was done by counting all the observations per individual station and checking whether that sum was smaller than 5% of the study period. The results of this step can be observed in figure 24 (bottom left graph).

5.2.3. Filtering out unrealistic measurements

The last step is primarily dedicated to filtering out measurements that are outliers, and are very highly unlikely to be accurate measurements. The initial method to detect outliers consist of selecting the mean residual value, minus/plus three times the standard deviation.

However, before that method can be used it is important to test if the data is normally distributed. If this would not be the case, and the data would for instance be heavily skewed, it would not make sense to have even thresholds on both sides of the mean. Nevertheless, the WOW-NL residuals seem to be relatively normally distributed, although it is slightly skewed as it shows more positive residuals (figure 24). Furthermore, the qq-plot indicates that the WOW-NL data is heavy tailed, which means that it includes more extreme values than should be expected from a normal distribution (figure 25). Applying the mean minus/plus three times the standard deviation method would mark all measurements that have a residual that is either bigger or smaller than 3.9°C as outliers, which comprises 1.24% of all the WOW-NL data. Since the WOW-NL residuals are not completely normally distributed, and using the mean minus/plus three times the standard deviation marks 1.24% as outliers, it is chosen to use a larger threshold. Accordingly, it is chosen to mark all the measurements that are 10°C above or below the predicted temperature as outliers. This results in a selection of 0.3%

of the data that are marked as outliers. Consequently, these will be removed from the WOW-NL dataset. When the WOW-NL data is assessed, and the gross errors an outliers are removed, the data integration part can start. The results from this part of the research are described in the following paragraph.

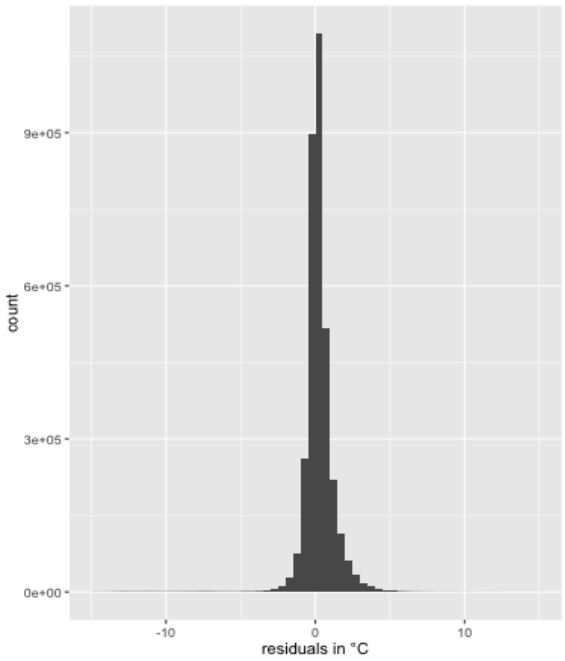


Figure 24: Histogram of the residuals from the WOW-NL observations versus the predicted temperatures.

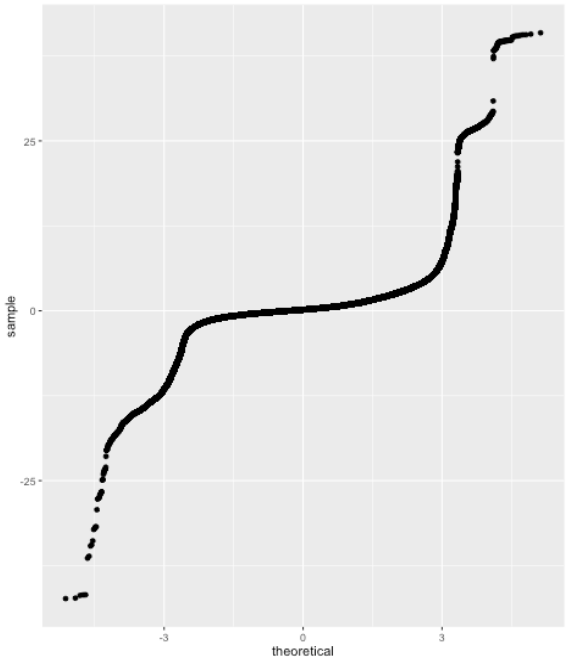


Figure 25: Normal Quantile-Quantile plot of the residuals from the WOW-NL observations versus the predicted temperatures.

5.3. Integration scenarios

The last part of this research consist of the data integration. This means that the WOW-NL data is integrated with the AWS data in order to explore its potential benefits. This was done according to three different scenarios that are characterised by a different methodology as was described in paragraph 4.3. Accordingly, the results that came from these different scenarios are described next.

5.3.1. WOW-NL as equal compared to AWS

The first integration scenario threats the WOW-NL data as if it is equal to the AWS data. Accordingly, there were no further modifications made to the WOW-NL data besides from the gross error and outlier removal. By using the leave-one-out cross validation it was tested whether the temperatures at the automatic weather stations could be predicted better using only AWS measurements, or by using the integrated data. This resulted in the following graphs (figure 26).

It becomes clear that there are many situations in which the WOW-NL data improve the interpolations and the results of the cross-validations. Although, it should be noted that in the summer the differences are not substantial. This is line with the other research results, as the WOW-NL observations deviate the most from the interpolations in the summer. October and January clearly show improvements in terms of RMSE compared to the interpolations based merely on automatic weather stations. Especially in January the improvement is considerable, as the RMSE of the WOWNL TPS interpolation stays much lower than the RMSE of the interpolation based on automatic weather stations. Besides that, the WOWNL IDW and the WOWNL OK interpolation method perform worse than the AWS interpolations. However, when comparing their performance with the results of the same method without the WOW-NL measurements, it becomes clear that they almost always perform better (table 3 & table 4).

	August	October	January
AWS IDW	0.96	0.80	1.24
AWS OK	0.86	0.71	0.93
AWS TPS	0.89	0.69	0.63

Table 3: Average RMSE values AWS

	August	October	January
WOWNL IDW	1.02	0.69	0.90
WOWNL OK	1.01	0.65	0.60
WOWNL TPS	0.97	0.65	0.42

Table 4: Average RMSE values WOWNL (AWS + WOW-NL)

Besides that, it should be stressed that the statistics also confirm that the integrated dataset performs worse overall for the summer. This is valid for all three interpolation methods. Accordingly, this suggests that the WOW-NL data is less accurate in the summer than in the other months, as they actually make the interpolation models perform worse, while they improve the sample size.

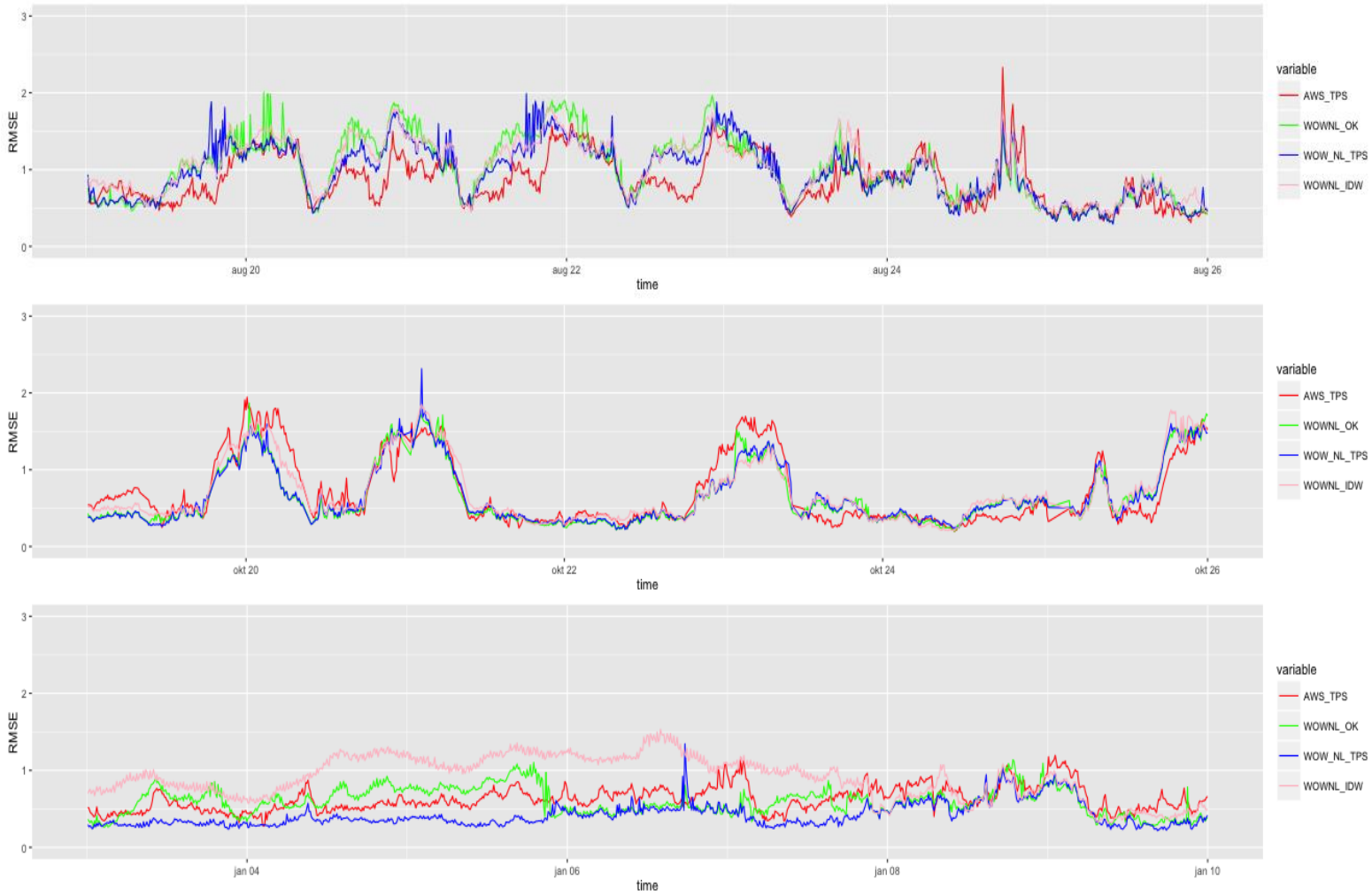


Figure 26: Results of the cross validation for the interpolation methods: AWS TPS (red), WOWNL OK (green), WOWNL TPS (blue), and WOWNL IDW (pink) for three different weeks in August, October, and January.

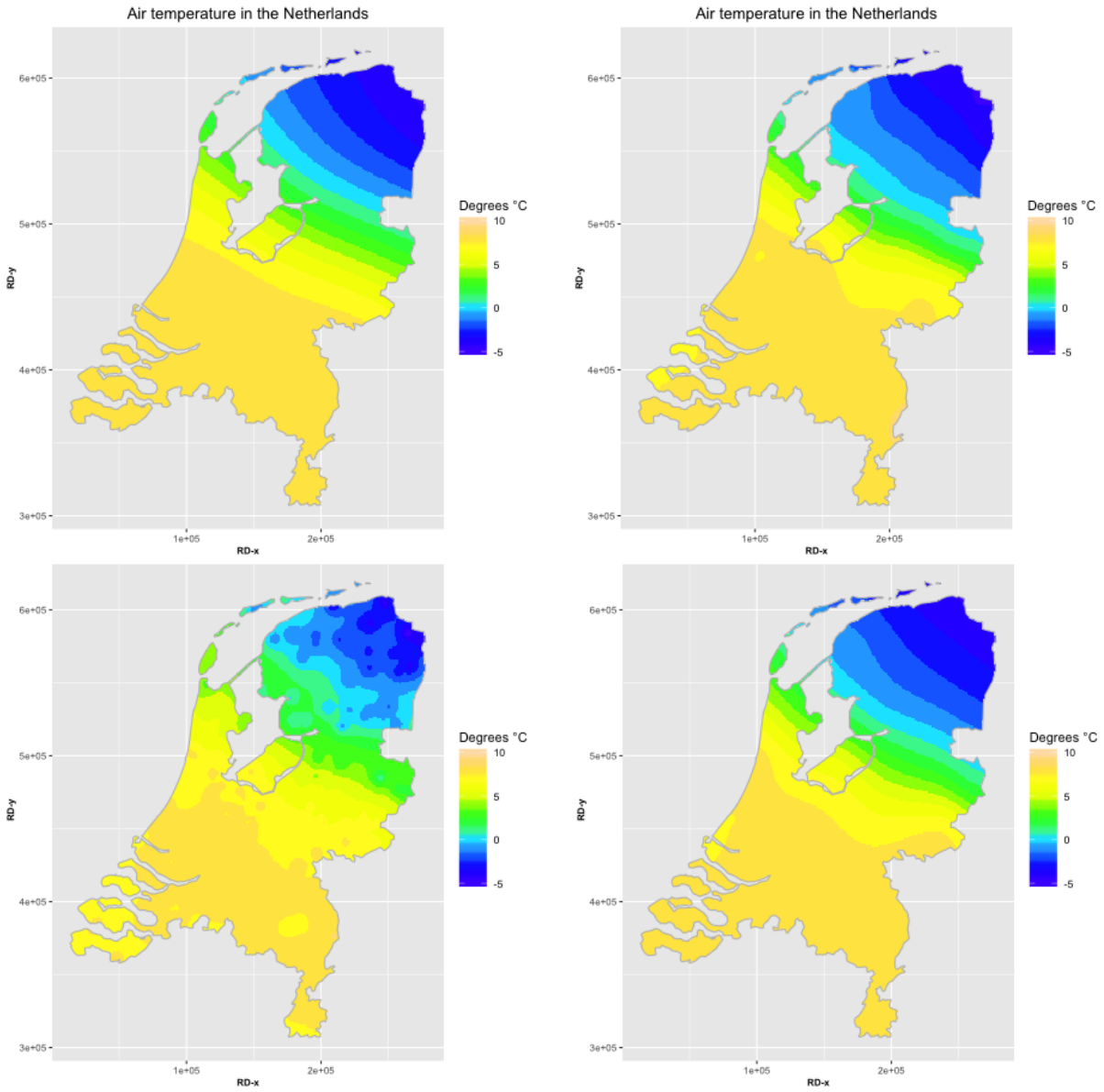


Figure 27: Air temperature on 05-01-2016 13:00 interpolated with different methods, from top left to bottom right: WOWNL OK, WOWNL TPS, WOWNL IDW, and AWS TPS.

Furthermore, the maps that are derived from the integrated dataset also show a clear enhancement over the map that is purely based on AWS measurements (figure 27). However, the WOWNL IDW interpolation produces an unrealistic map which includes an omnipresent circle pattern. The WOWNL OK and the WOWNL TPS interpolation produce smooth maps that illustrate data with an improved spatial resolution. Furthermore, it should be noted that this is a situation in which the interpolations with the integrated data have a low RMSE. There are also situations where this is not the case.

Overall, it should be concluded that the data integration can in some situation lead to improved temperature interpolations, both in terms of performance indicators as well as visually. Although, most of the time in August, the integrated dataset made interpolations deteriorate in terms of RMSE. This suggests that the WOW-NL data has less predictive power in the summer, as normally an increase in observation sample size should improve interpolation quality. Nevertheless, October and January showed that the integration of the WOW-NL data improved the interpolations.

5.3.2. WOW-NL as secondary predictive variable

The second integration scenario threats the WOW-NL data as less reliable than the AWS data. This is done by only using the WOW-NL data as a secondary predictive variable in the UK interpolation method (as was described in paragraph 4.3.2). However, the first step comprises an interpolation of the WOW-NL observations for a continuous grid, in order to have temperature estimates on the locations of automatic weather stations (figure 28 shows one example). This is done for every time step so that each AWS measurement has both an observed temperature and an interpolated WOW-NL temperature. This procedure made it possible to use the WOW-NL data as a secondary variable in the UK interpolation method.

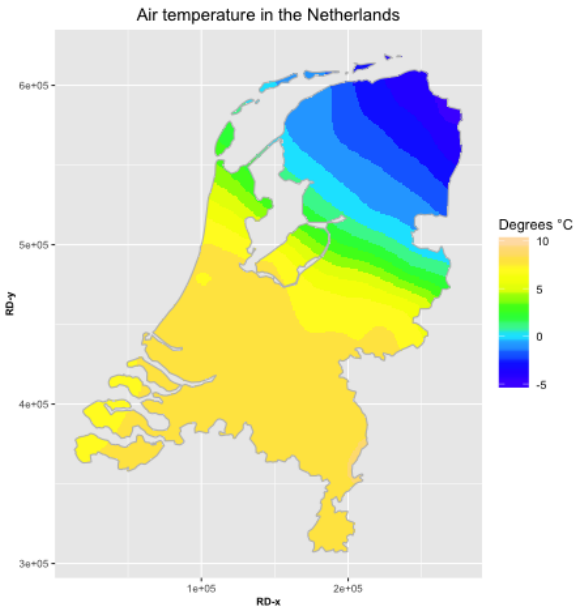


Figure 28: Air temperature on 05-01-2016 15:30 interpolated with WOWNL TPS

WOW-NL Universal Kriging (UK) interpolation results

The second step of this scenario consists of making interpolations and cross-validations for each time step during the same study period as was used during the previous scenario. The results of the cross-validation give a comprehensive overview of how successful the WOWNL UK interpolation is through time (figure 29). Accordingly, it becomes clear that during August, the WOWNL UK interpolation method does not seem to improve the interpolation results much. Most of the time, the AWS TPS interpolation is quite similar in terms of RMSE. However, it should be mentioned that there are also some time steps where either the WOWNL UK or the AWS TPS interpolation performs worse. Besides that, in August the WOWNL UK performs slightly better overall with an RMSE of 0.81 compared to 0.89 for AWS TPS (table 5).

Furthermore, in October the WOWNL UK consistently scores much better in terms of RMSE than the AWS TPS interpolation. This improvement is also confirmed by the average RMSE, which is 0.56 for WOWNL UK compared to 0.69 for AWS TPS. It should therefore be concluded that using the WOW-NL data as a secondary predictive variable in October improves the quality of the interpolations.

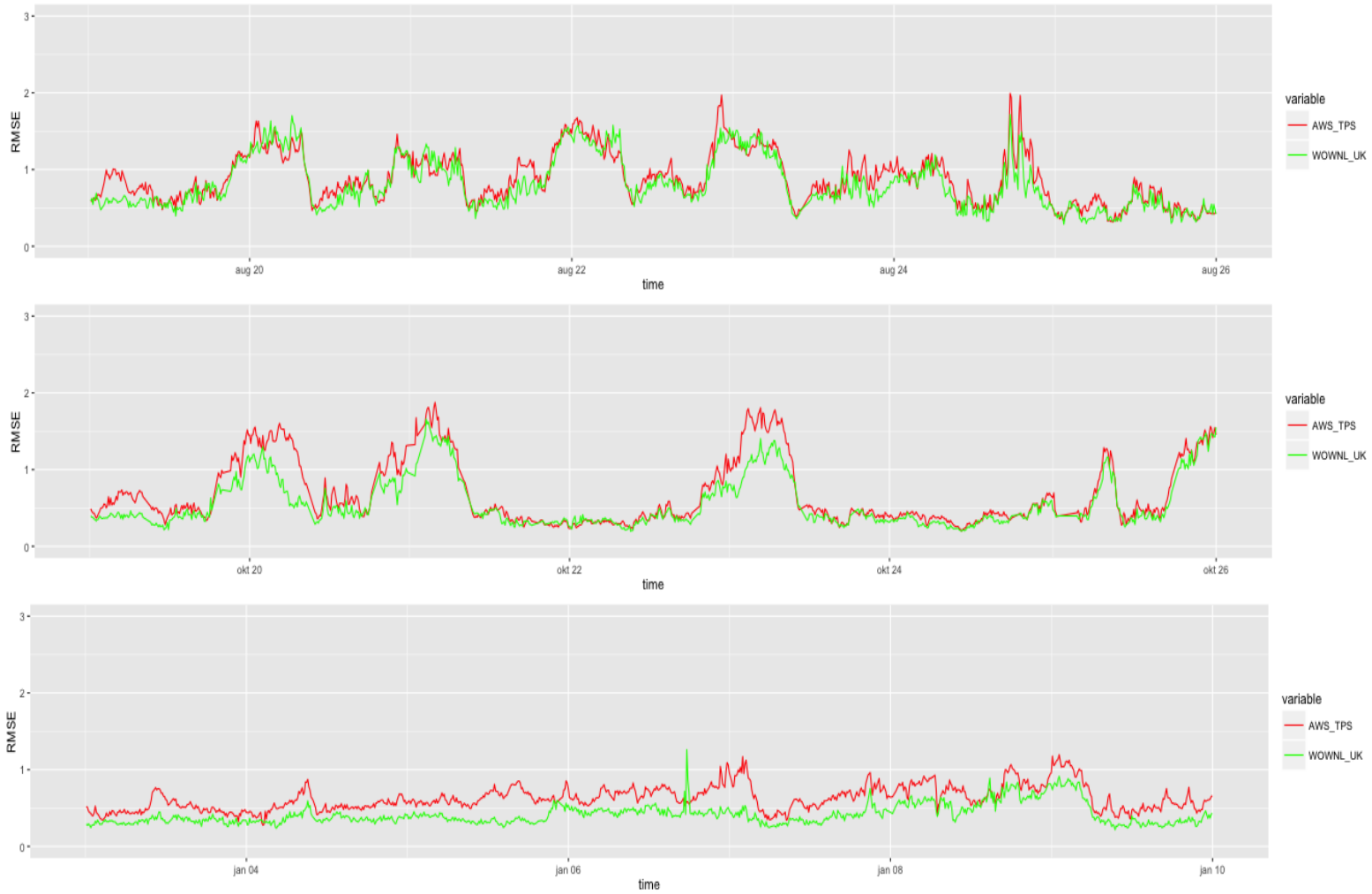


Figure 29: Results of the cross validation in RMSE of the AWS TPS (red) interpolation compared with the WOWNL UK (green) interpolation for three different weeks in August, October, and January.

This is also the case for January, where the WOWNL UK deviates the most from the AWS TPS interpolation. Also, the WOWNL UK interpolation structurally outperforms AWS TPS in terms of RMSE. Equally, it should be mentioned that there are only a few exceptions where the RMSE of the WOWNL UK is higher than the AWS TPS. The average RMSE of the WOWNL UK is 0.42 compared to 0.63 for AWS TPS.

	August	October	January
AWS_OK	0.86	0.71	0.93
AWS_TPS	0.89	0.69	0.63
WOWNL_UK	0.81	0.56	0.42

Table 5: Average RMSE values interpolation methods

Besides the performance of the WOWNL UK interpolation in terms of RMSE, it is also important to examine the results visually (figure 30). It should be noted that the WOWNL UK interpolation result is also dependent on the interpolations that were done in order to generate a continuous surface based on WOW-NL measurements (figure 28). Since this was done using the TPS method, the resultant UK interpolation includes TPS patterns. However, between the three different interpolations there is quite some difference. It is important to stress that the WOWNL UK interpolation performs much better in this particular example (figure 30).

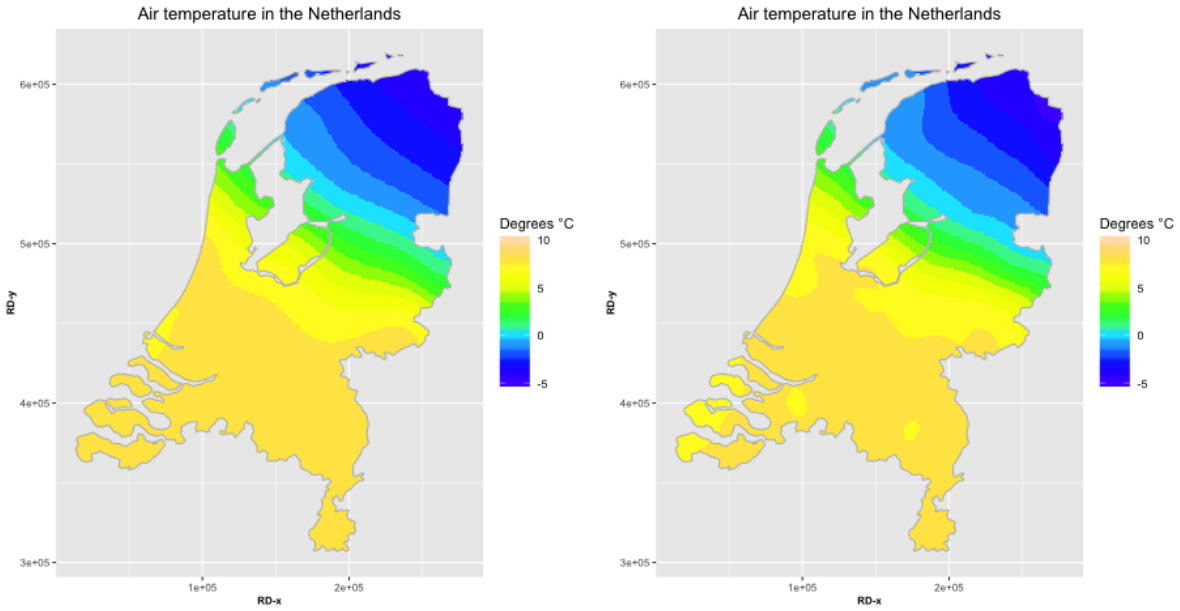


Figure 30: Air temperature on 05-01-2016 15:30 interpolated with different methods, from left to right: AWS TPS and WOWNL UK.

Finally, it should be concluded that using WOW-NL data as a secondary variable can improve temperature interpolations for the Netherlands. Although, it should be mentioned that during August the difference was negligible. For October and January the differences seem more relevant as they are much bigger. This is also in alignment with the findings from Bell (2014), the results from the data assessment, and integration scenario 1. Although it cannot unequivocally be proven that the WOW-NL measurements are worse in summer, the indirect based evidence all points in this direction. Nevertheless, it should be stressed that these findings are based a cross-validation with automatic weather stations. Accordingly, it is impossible to determine if interpolations improve for all places in the Netherlands. Conversely, it is certain that the WOW-NL as secondary variable improves temperature estimations for the sites of automatic weather stations.

5.3.3. Correcting WOW-NL observations for solar radiation

The last scenario aims to further correct the WOW-NL data before it is actually integrated. Accordingly, a multiple regression analysis and a random forest analysis were conducted in order to examine the relation between the interpolated temperature, incoming solar radiation, the time of measurement, and the observed WOW-NL temperatures. The first step consisted of an interpolation of radiation measurements by the automatic weather stations as was described in paragraph 4.3.3. As a result, each WOW-NL measurement included a prediction of the average incoming sunlight. This step made it possible to add incoming solar radiation in the prediction models.

Multiple regression analysis for all data

The multiple regression analysis shows that solar radiation has a significant positive relation with the predicted temperature (figure 31). However, it should be noted that the predictive power is not as high as was expected. The most important element in the multiple regression analysis consist of the observed temperatures at the WOW-NL stations.

Formula:
Predicted temperature ~ observed temperature + time step + radiation

Residuals:

Min	1Q	Median	3Q	Max
-9.6940	-0.3520	0.0818	0.4528	9.8451

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.925e+00	1.953e-01	-25.21	<2e-16 ***
temperature	9.620e-01	1.153e-04	8343.13	<2e-16 ***
time step	3.499e-09	1.344e-10	26.04	<2e-16 ***
radiation	3.634e-04	9.211e-06	39.46	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8405 on 3330056 degrees of freedom
Multiple R-squared: 0.9814, Adjusted R-squared: 0.9814
F-statistic: 5.863e+07 on 3 and 3330056 DF, p-value: < 2.2e-16

Figure 31: Summary report of the multiple regression analysis.

Altogether, the model has a relatively high R-squared, which means that a considerable share of the variance is explained by the model (98%). Nevertheless, it should also be stressed that the model does not fulfil all the necessary requirements, as the residuals are not normally distributed. Furthermore, it should also be assessed: if the other variables are normally distributed, if there is multicollinearity, if there is no auto-correlation, and if there is no homoscedasticity (Montgomery et al., 2015). Since the objective is not to make a generic statistical model, and the predictive power of the radiation is relatively low, it is chosen not to go in further detail about the remaining statistical assumptions.

Furthermore, the model was used to make corrections to all the WOW-NL data. As is shown in figure 32, the changes are relatively small. Since the multiple regression model is determined by the best fit based on the smallest sum of the residuals, it does not make drastic changes. The total RMSE of the

predicted temperature versus the observed temperature changes from 1.32, to 1.24. Accordingly, this should be acknowledged as only a small improvement.

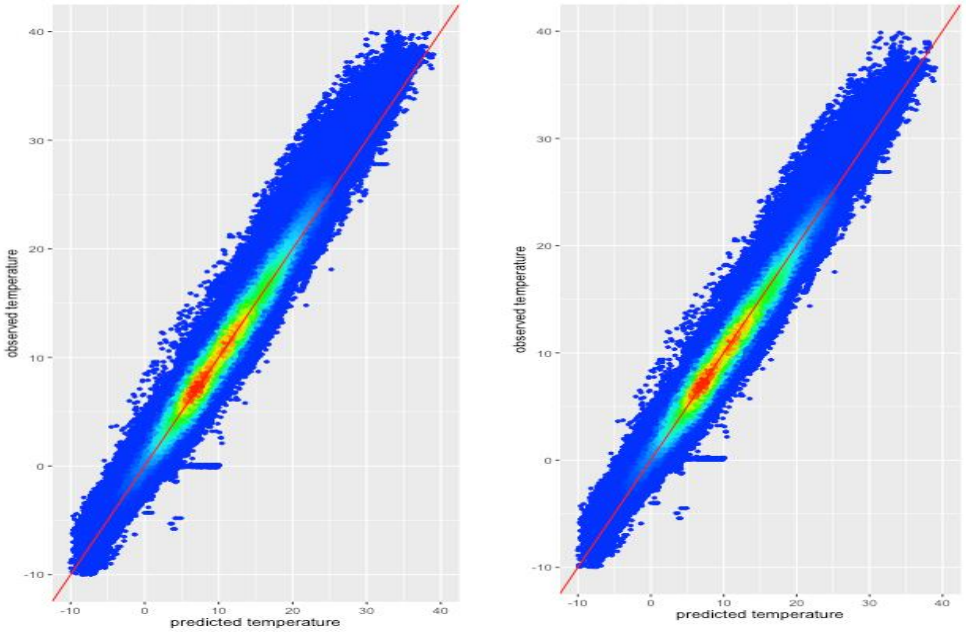


Figure 32: Predicted temperatures plotted versus the observed temperatures by WOW-NL stations. From left to right: before correction model, and after correction model. Color scale shows density from blue (least dense) to red (most dense).

Random forest model for hot days

When using the random forest model for two of the warmest days in August, the correction model seems to perform substantially better. Where the multiple regression analysis is restricted to a linear fit, the random forest model is not. As a result, the random forest model is capable of making more drastic corrections than the multiple regression analysis (figure 33). However, it should be noted that these two exemplary hot days are easier to model than the whole dataset. Nevertheless, the RMSE of the predicted temperature versus the observed temperature improves from 1.58 to 0.66. Besides that, it becomes clear that during these hot days, the incoming solar ration has much more predictive power compared to the multiple regression analysis for the whole dataset (figure 34). Additionally, figure 34 shows that the random forest model levels out in terms of mean error after approximately 50 trees.

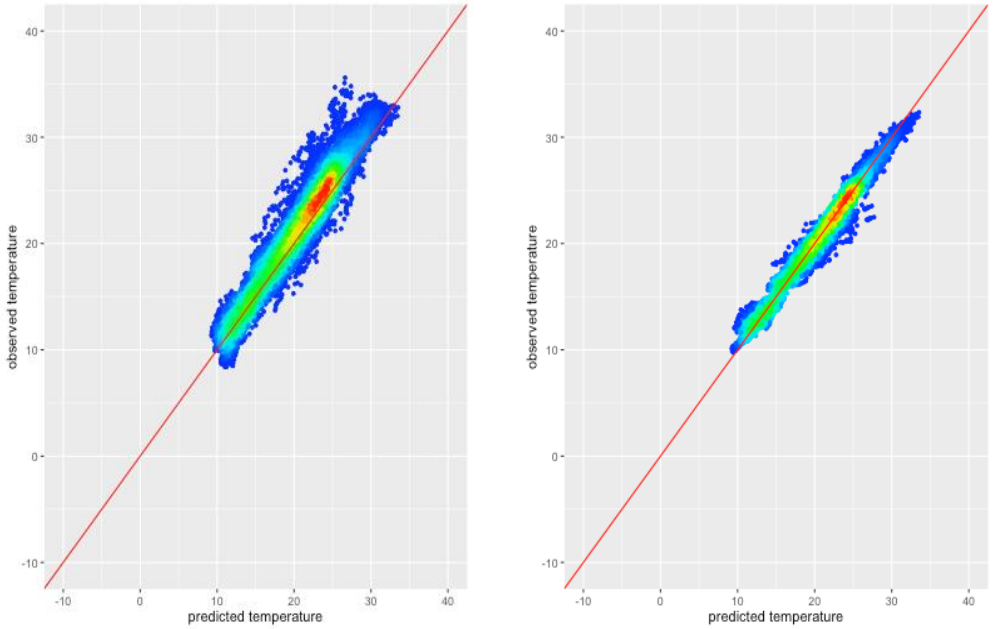
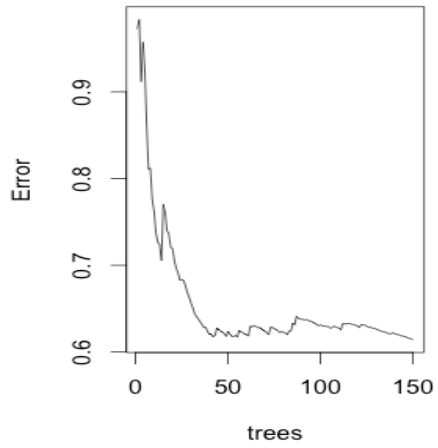


Figure 33: Predicted temperatures plotted versus the observed temperatures (2016-08-05 till 2016-08-07) by WOW-NL stations. From left to right: before random forest model, and after random forest model. Color scale shows density from blue (least dense) to red (most dense).



Variable	Variable importance (% Increase MSE)
Radiation	19.6
Observed temperature	21.2
Time step	23.6

Figure 34: The amount of trees plotted versus the Mean error of the model and its relative variable importance.

Furthermore, it is also important to examine the results of the correction models visually, by making maps with the corrected data (figure 35). Accordingly, it becomes clear that the WOW-NL TPS map, which is made with the corrected data from the random forest model, shows the most resemblance with the AWS TPS map. This is according to expectations, since the model performs well and takes the interpolated temperature as the dependent variable. However, there are still some differences with the AWS TPS map, which are the result of the unexplained variance in the model. Furthermore, the multiple regression model made less powerful corrections, hence the map that is based on its corrections deviates most from the AWS TPS map. Since the corrections are both based on models that take the interpolated temperature as the independent variable, it does not make sense to test the quality with a cross-validation. This is due to the fact that the interpolations are based on the AWS measurements, and the cross-validation can only be done with AWS measurements.

Finally, it should be argued that both models showed that incoming solar radiation was significantly related to interpolated temperatures. However, the predictive power was less than was expected in the multiple regression analysis. The random forest model for two exemplary hot days in August showed that solar radiation has more predictive power for warm days. Overall, the corrections by the random forest model were more drastic than the regression analysis. Finally, maps made with the corrected data show that although corrections have been made, there is still substantial difference from the AWS interpolations. However, it should be acknowledged as a limitation that the quality of the interpolations cannot be tested according to a cross-validation.

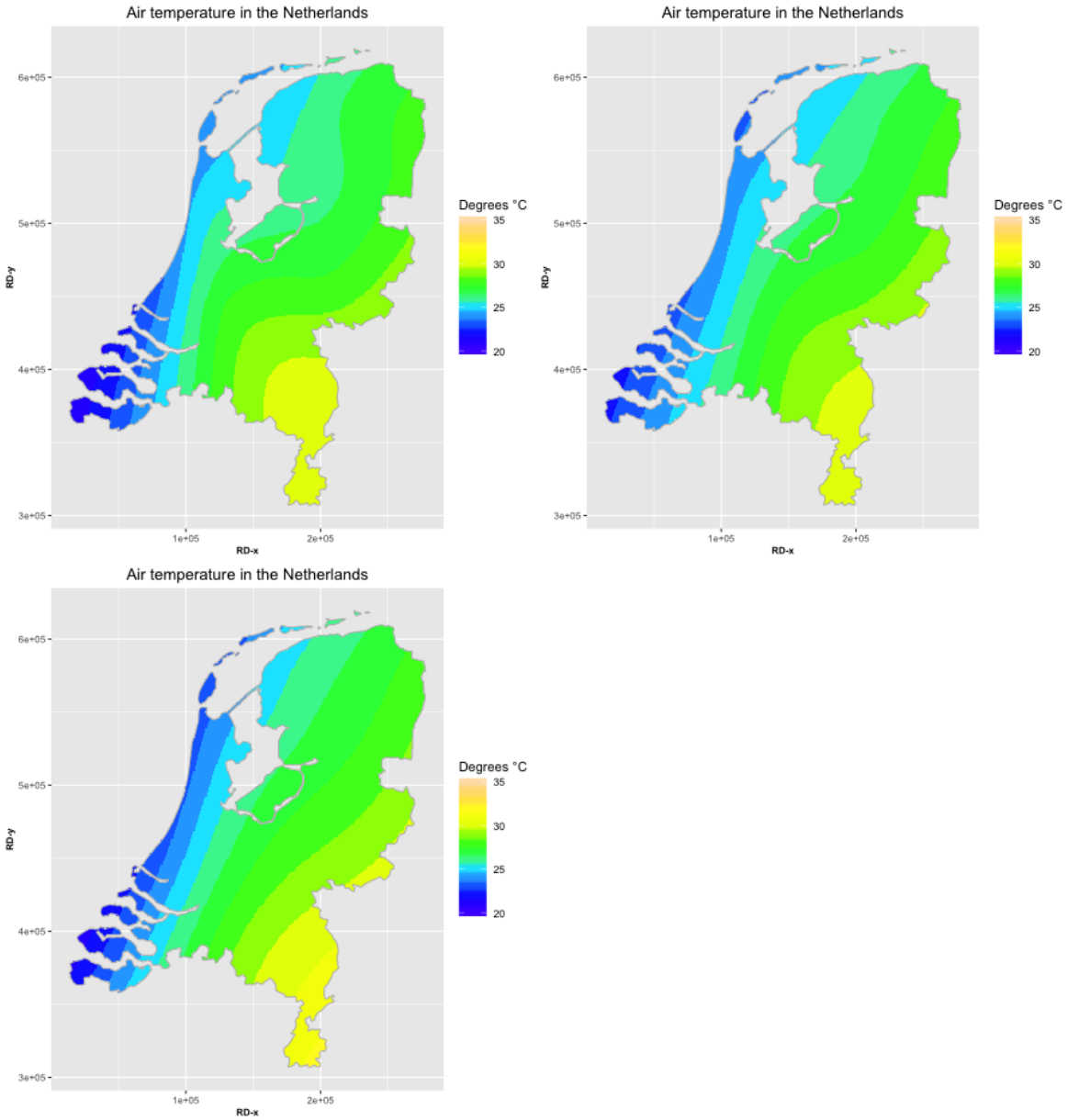


Figure 35: Air temperature on 06-08-2015 13:00 interpolated with different methods. From top left to bottom right: WOWNL TPS after corrections with multiple regression model, WOWNL TPS after corrections with random forest model, and AWS TPS.

6. Discussion

The objective of this chapter is to reflect on the overall research results, the methods that were used, and the extent to which the findings are in line with other research. Accordingly, the most important findings of this research are discussed.

Firstly, it should be noted that determining the quality of the WOW-NL data turned out to be a rather complex task, which could only partly be done. This is due to the fact that it was chosen to compare WOW-NL observations with reliable temperature estimates (AWS interpolations). When a temperature estimate deviates from an WOW-NL observations, it remains virtually impossible to assess which part consists of all the factors that together comprise the bias of a WOW-NL station (e.g., radiation bias, calibration bias, or representativity bias), and which part consists of the actual variation in temperature that exists across space. Nevertheless, this method does illuminate which observations are most likely gross errors and outliers. Besides that, the data assessment also showed that the WOW-NL observations deviate most from the temperature estimates during the daytime in summer. This is in line with previous research as Bell (2014) argues that incoming solar radiation is one of the most challenging obstacles for amateur weather stations.

Furthermore, the gross error and outlier removal showed that only 1.24% of all WOW-NL observations deviate more than 3.9°C from the temperature estimates. Although this includes thousands of measurements, it is only a small part of the WOW-NL data. As a result, it should be concluded that potential quality control methods do not have to remove much data, and do not have to be very complex.

The data integration showed that various methods could be used to exploit the potential benefits that are associated with the integration of WOW-NL data and formal KNMI data. From the three methods that were used, the second integration scenario showed the most promising results, as the average interpolation quality in terms of RMSE improved for the whole study period.

However, for the first two scenarios it should be noted that it remains hard to tell whether the interpolations improve for all places in the Netherlands, since it can only be tested for the locations of the automatic weather stations. In addition, it should be noted that the data assessment merely illuminated that WOW-NL observations deviate from the interpolations most is August. This does not necessarily mean that these observations are wrong. However, the results from the first integration scenario also show that the quality of temperature interpolations in August decreases with the integration of WOW-NL data. Although it cannot unequivocally be proven, this strongly suggests that the WOW-NL measurements are also less accurate in summer. Equally, this is most likely caused by the former mentioned radiation bias.

The third integration scenario showed that correction could be made for incoming solar radiation. However, since the corrections are both based on models that take the interpolated temperature as the independent variable, it does not make sense to test the quality with a cross-validation. This makes it impossible to test whether the corrected data has actually improved the interpolations. As a result, this should be regarded as an important limitation.

Furthermore, in this research temperature was studied on the scale at which the KNMI measures, while many WOW-NL users might be interested in micro climates (e.g., gardens, urban areas). Accordingly, these WOW-NL stations can also provide useful data, although their quality should not be assessed with the methods that were used in this research.

7. Conclusions

The most important objective of this chapter is to adequately answer the main research question, which was formulated as follows:

> ***"To what extent can the integration of spatial data, derived from amateur weather stations in the WOW-NL application, and formal KNMI data improve the spatiotemporal resolution of meteorological data?"***

Furthermore, the main research question was answered according to a set of sub-questions. These are all answered in order to gain a comprehensive overview of the most important conclusions regarding this research. Accordingly, the sub-questions were formulate as follows.

1) *What are the benefits and shortcomings of Volunteered Geographic Information that is derived from sensors?*
2) *How can the quality of data derived from amateur weather stations be determined?*
3) *How accurate is the data that is derived from amateur weather stations in the WOW-NL application?*
4) *How can the data that is derived from amateur weather stations in the WOW-NL application improve the spatiotemporal resolution of temperature data and maps produced by the KNMI?*

1). Firstly, it became clear that VGI is a source of spatial data with much potential. However, the use of VGI also requires caution since there is a considerable lack of quality control and validation standards. As a result, substantial pre-processing is necessary in order to filter out possible corrupted measurements. Besides the general limitations, there are also some case specific issues that apply especially to data from amateur weather stations. These include: calibration issues, design flaws, communication and software errors, metadata issues, and representativity errors. Given the former mentioned issues, it is concluded that quality control is imperative for the utilization of the WOW-NL data.

2). In order to assess the quality of WOW-NL data, its measurements were compared with reliable temperature estimates. Accordingly, these were derived from interpolations based on formal AWS data. However, this required the selection of the best interpolation method for temperature measurements with a 10 minute temporal resolution. An extensive inter-comparison showed that TPS is the best performing interpolation method in this regard.

3). The comparison of the WOW-NL observations and the interpolations showed that the WOW-NL stations generally observe higher temperatures than the interpolated temperatures. Examining the individual stations showed that some stations are very closely following the interpolated temperatures, while other stations only showed big residuals when the interpolated temperatures went above approximately 20 °C. Besides that, there were also stations that were clearly registering outliers and errors. Considering the temporal subsets, it should be concluded that the observed WOW-NL observations deviated the most from the interpolations during the summer (daytime). It is strongly presumed that this due to the relatively high solar radiation during the summer since it is known that this can significantly deteriorate the quality of the observations from amateur weather stations. Furthermore, the station attributes (station rating, ucz index, exposure rating, and temperature instrument) did not show notable differences among the various subsets. In addition, the stations from reputable brands did not show closer resemblance to the interpolated temperatures either.

4). Before the WOW-NL data could be integrated with the formal AWS data, it was necessary to conduct a gross error and outlier removal. Accordingly, the following type of WOW-NL measurements were removed: repetitive measurements, measurements from stations that were off most of the time, and unrealistic measurements. Removing the WOW-NL observations that were 10 °C higher or lower that than the interpolated temperature proved to be a suitable method of filtering out unrealistic measurements.

Finally, the formal KNMI data and the WOW-NL data were integrated according to three different scenarios. The first scenario threats the WOW-NL data as if it is equal to AWS data. The results of the first scenario showed that the integrated data improved the temperature interpolations for the Netherlands in both October and January. However, in August the integrated data did not improve the interpolations. This is most likely due to radiation bias, which deteriorates the quality of WOW-NL observations in summer.

In the second scenario the WOW-NL data was treated as less reliable than the AWS data. This was done by using the WOW-NL data only as a secondary predictive variable in the UK interpolation method. Equally, the integrated data improved interpolations drastically for October and January. For August, the improvements were negligible.

Furthermore, the third scenario aimed to correct the WOW-NL data before the integration. Accordingly, a multiple regression analysis showed that corrections could be made according to the time step, the observed WOW-NL temperature, and incoming solar radiation. However, the corrections only resulted in a marginal improvement over the total original WOW-NL data. When the same relation was modeled with a random forest model for exemplary warm days in August, more drastic corrections could be made, and solar radiation had more predictive power.

Finally, it should be concluded that the integration of the WOW-NL data and AWS data can improve the spatiotemporal resolution of meteorological data and maps. However, the extent to which this is true is highly dependable on the time of the year and the data integration method. Furthermore, the selection of the interpolation method can also have considerable influence on the quality improvements associated with the data integration.

8. Recommendations

The final chapter of this master thesis includes recommendations for the KNMI. As the WOW-NL project continues after the finalisation of this research, recommendations can both contribute to policy and further research. Accordingly, the aim of this recommendation is to include helpful suggestions for both domains.

Regarding the policy of the WOW-NL applications, different areas of potential improvement became apparent during this research. One of the biggest limitations regarding the WOW-NL data included the lack of metadata. Accordingly, it is known that metadata is one of the most important elements that enables efficient use of spatial data. This especially true when a spatial database includes data from heterogeneous sources, which is also the case for the WOW-NL database. Besides that, using metadata enables users of spatial data to assess whether data aligns with their specific needs in terms of usability and quality. It would for instance be incredibly useful to know which type of weather station (model) is responsible for which observations. This would give users of the WOW-NL data better insight regarding the quality of the measurements. Additionally, this would also enhance the data assessment for the KNMI, since similar weather stations will generally have the same type of bias. Consequently, it is advised to examine the possibilities to include the station model in the required information that owners of a weather station have to disclose when connecting with the WOW-NL application.

Furthermore, it became clear that determining the representativity of WOW-NL stations is a relatively complex task, despite of the sophisticated site rating criteria that are available in the metadata. However, if the owners of WOW-NL stations would have the possibility to include a digital image of their specific set-up, it would become easier to assess the representativity of a WOW-NL station. Consequently, it is advised to examine the possibility to add this functionality to the WOW-NL application. However, it should be stressed that this requires more effort from the volunteer, hence it should not be a mandatory request. A good example of this possibility is already demonstrated by the WOW-UK application.

During the meetings with the project team, it became apparent that there is no or little communication between the KNMI and the owners of WOW-NL stations. This is partly due to technical restrictions, which withholds the personal information of the WOW-NL station owners from the KNMI. However, it is strongly advised to overcome this restriction, and start communications in order to solidify the WOW-NL community. Additionally, contact with the owners of WOW-NL stations can potentially result in some practical benefits as well. Whenever, a WOW-NL station shows suspicious measurements (e.g., repetitive, unrealistic, or no measurements), it would be possible to warn the owners in order to solve the issues. Additionally, it would become possible to give more general (automatic) feedback regarding the quality of the measurements. As a result, the quality of the WOW-NL data can be improved and additionally, a more lively community can be created.

Besides that, this research illuminated that the WOW-NL data includes a substantial amount of gross errors and outliers. However, after the removal of these erroneous measurements, the WOW-NL data turns out to have much potential. It should be acknowledged as beneficial if such measurements could be removed a priori instead of a posteriori. Hence, it is advised to examine the possibilities to construct a quality control method, which is capable of safeguarding the database from the most common errors before they are stored.

Besides the various recommendations that aim to contribute to the policy regarding the WOW-NL application, there are also various suggestion regarding further research. Firstly, it should be noted that this research only covered one variable (temperature), while most WOW-NL stations measure several others (e.g., precipitation, wind, and air pressure). Since the results of this research illustrated that the WOW-NL temperature data includes the potential to enhance the spatiotemporal resolution of formal data, it is worth testing to what extent this is valid for other variables.

Furthermore, this research primarily focussed on temperature at the scale at which the KNMI measures. However, the WOW-NL data potentially contains data that can be used to study temperature at different scales (e.g., urban heat island) or even smaller micro climates (e.g., villages, neighborhoods, streets, etc.). Hence, it would be useful to examine relations between the WOW-NL data, and other spatial data sources that describe the environment in which the WOW-NL stations are placed. Other related sources could for instance include: land use data, elevation models, satellite data, etc. Accordingly, further research could illuminate to what extent the WOW-NL data can be used to study temperature at smaller scales.

Besides that, the WOW-NL data can potentially also be used to study extreme situations. This research primarily studied the overall potential of WOW-NL temperature data in everyday situations. However, extreme weather can have much local variety that is hard to capture with only the AWS stations. Accordingly, the WOW-NL data can potentially be used as a useful source that sheds light on local situations in extreme weather conditions.

Concluding, it should be stressed that the KNMI showed considerable interest in WOW-NL project and the data that is derived from the WOW-NL application. The commissioned researches of Koolen (2016), and this master thesis illustrate this. As a result, it is therefore advised to conserve this policy, and to continue with data assessment and research regarding the improvement and usability of WOW-NL data.

References

Bell, S., 2014. *Quantifying Uncertainty in Citizen Weather Data*. Aston University.

Bell, S., Cornford, D. & Bastin, L., 2013. The state of automated amateur weather observations. *Weather*, 68(2), pp.36–41. Available at: http://doi.wiley.com/10.1002/wea.2044.

Blumer, A. et al., 1990. Occam's razor. *Readings in machine learning*, 201-204, pp.377–380.

Breiman, L., 1999. Random Forests. *Machine Learning*, 45(5), pp.1–35.

Burke, S., 2001. Missing Values, Outliers, Robust Statistics & Non-parametric Methods. , pp.19–24.

Castell, N. et al., 2014. Mobile technologies and services for environmental monitoring : The Citi-Sense-MOB approach. *Urban Climate*, (November 2015).

Chapman, L. et al., 2014. Winter Road Maintenance and the Internet of Things. , pp.1–7.

Cinnamon, J., 2015. Deconstructing the binaries of spatial data production: Towards hybridity. *The Canadian Geographer / Le Géographe canadien*, 59(1), pp.35–51.

Corke, P. et al., 2010. Environmental wireless sensor networks. *Proceedings of the IEEE*, 98(11), pp.1903–1917.

Dirksen, M., 2015. High Resolution Temperature Interpolation : combining observations and model data. , (October).

Dobesch, H., Dumolard, P. & Dyras, I., 2007. *Spatial Interpolation for Climate Data: The Use of GIS in Climatology and Meteorology*, Available at: http://doi.wiley.com/10.1002/9780470612262.

Douglas, N. et al., 2016. Fields. Available at: https://cran.r-project.org/web/packages/fields/index.html.

Dyras, I. et al., 2005. *The use of Geographic Information Systems in climatology and meteorology: COST 719*, Available at: http://doi.wiley.com/10.1017/S1350482705001544.

Einstein, A., 2010. *The ultimate quotable Einstein*, Princeton University Press.

ESRI, 2015a. How Inverse Distance Weighted (IDW) interpolation works. Available at: http://support.esri.com/.

ESRI, 2015b. How Kriging works. Available at: http://support.esri.com/.

ESRI, 2015c. Spatial autocorrelation. Available at: http://support.esri.com/.

ESRI, 2015d. Understanding geostatistical analysis. Available at: http://support.esri.com/.

Feick, R. & Roche, S., 2013. Understanding the Value of VGI. In S. Daniel, E. Sarah, & M. Goodchild, eds. *Crowdsourcing Geographic Knowledge*. pp. 15–29. Available at: http://link.springer.com/10.1007/978-94-007-4587-2.

Flanagin, A.J. & Metzger, M.J., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4), pp.137–148. Available at: http://link.springer.com/10.1007/s10708-008-9188-y [Accessed July 12, 2014].

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221. Available at: http://link.springer.com/10.1007/s10708-007-9111-y [Accessed July 14, 2014].

Haklay, M., 2013. Crowdsourcing Geographic Knowledge. , pp.105–122. Available at: http://link.springer.com/10.1007/978-94-007-4587-2.

Hengl, T., 2007. *A Practical Guide to Geostatistical Mapping og Environmental Variables*,

Hiemstra, P., 2015. automap. Available at: https://cran.r-project.org/web/packages/automap/index.html.

Hiemstra, P. & Sluiter, R., 2011. Interpolation of Makkink evaporation in the Netherlands. , p.78. Available at: http://www.numbertheory.nl/files/report_evap.pdf.

Hofstra, N. et al., 2008. Comparison of six methods for the interpolation of daily, European climate data. *Journal of Geophysical Research*, 113(D21), p.D21110. Available at: http://doi.wiley.com/10.1029/2008JD010100.

Joly, D. et al., 2011. Temperature interpolation based on local information: the example of France. *International Journal of Climatology*, 31(14), pp.2141–2153. Available at: http://doi.wiley.com/10.1002/joc.2220 [Accessed December 1, 2015].

Koninklijk Nederlands Meteorologisch Instituut, 2016a. Uurgegevens van het weer in Nederland. Available at: http://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.

Koninklijk Nederlands Meteorologisch Instituut, 2016b. WOW-NL jouw weer op de kaart. Available at: https://wow.knmi.nl/.

Koolen, M., 2016. *Rapportage WOW-NL evaluatie: Vergelijking WOW met KNMI waarnemingen (To be published)*, De Bilt.

Lam, N.S.-N., 1983. Spatial Interpolation Methods: A Review. *The American Cartographer*, 10(2), pp.129–150.

Leys, C. et al., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), pp.764–766.

Li, J. & Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6.3, pp.228–241.

Li, J. & Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. *Australian Geological Survey Organisation*, GeoCat# 68(2008/23), p.154.

Li, J. & Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53, pp.173–189. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1574954110001147.

Liaw, a & Wiener, M., 2002. Classification and Regression by randomForest. *R news*, 2(December), pp.18–22.

Meng, Q., Liu, Z. & Borders, B.E., 2013. Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods. *Cartography and Geographic Information Science*, 40(1), pp.28–39. Available at: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.762138.

Meteorological Office, 2016. Weather Observation Website. Available at: http://wow.metoffice.gov.uk/.

Montgomery, D.C., Peck, E.A. & Vining, G.G., 2015. *Introduction to linear regression analysis.*, Chicago: John Wiley & Sons.

Muller, C.L. et al., 2015. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, (JANUARY), p.n/a–n/a. Available at: http://doi.wiley.com/10.1002/joc.4210.

Oliver, M.A. & Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena*, 113, pp.56–69. Available at:

http://linkinghub.elsevier.com/retrieve/pii/S0341816213002385.

Osborne, J.W. & Overton, A., 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(1976), p.Retrieved June 1, 2012 from http://PAREonline.net/.

Pebesma, E. & Graeler, B., 2016. gstat. Available at: https://cran.r-project.org/web/packages/gstat/index.html.

Salet, F.W.J., 2009. Het interpoleren van temperatuurgegevens. *Koninklijk Nederlands Meteorologisch Instituut*. Available at: http://bibliotheek.knmi.nl/stageverslagen/stageverslag_Salet.pdf.

Santamouris, M., 2014. Cooling the cities - A review of reflective and green roof mitigation technologies to fight heat island and improve comfort in urban environments. *Solar Energy*, 103, pp.682–703. Available at: http://dx.doi.org/10.1016/j.solener.2012.07.003.

Sluiter, R., 2009. Interpolation methods for climate data: literature review. *Koninklijk Nederlands Meteorologisch Instituut*. Available at: http://bibliotheek.knmi.nl/knmipubIR/IR2009-04.pdf.

Sluiter, R., 2012. Interpolation Methods for the Climate Atlas. *Koninklijk Nederlands Meteorologisch Instituut*. Available at: http://bibliotheek.knmi.nl/knmipubTR/TR335.pdf.

Sosko, S. & Dalyot, S., 2015. Towards the Use of Crowdsourced Volunteered Meteorological Data for Forest Fire Monitoring. , (c), pp.127–132.

The Times, 2011. Experts make way for an outbreak of enthusiasm from citizen forecasters. , pp.1–3.

Tomczak, M., 1998. Spatial Interpolation and its Uncertainty Using Automated Anisotropic Inverse Distance Weighting (IDW) - Cross-Validation/Jackknife Approach. *Journal of Geographic Information and Decision …*, 2(December), pp.18–30.

Williams, M. et al., 2011. Automatic processing, quality assurance and serving of real-time weather data. *Computers & Geosciences*, 37(3), pp.353–362. Available at: http://dx.doi.org/10.1016/j.cageo.2010.05.010.

Willmott, C.J. & Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), pp.79–82.

Wolters, D. & Brandsma, T., 2012. Estimating the Urban Heat Island in Residential Areas in the Netherlands Using Observations by Weather Amateurs. *Journal of Applied Meteorology and Climatology*, 51(4), pp.711–721. Available at: http://journals.ametsoc.org/doi/abs/10.1175/JAMC-D-11-0135.1.

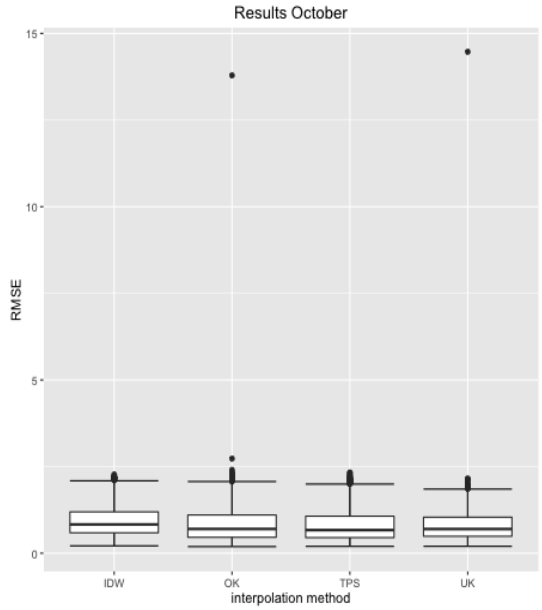
Wu, T. & Li, Y., 2013. Spatial interpolation of temperature in the United States using residual kriging. *Applied Geography*, 44(OCTOBER 2013), pp.112–120. Available at: http://dx.doi.org/10.1016/j.apgeog.2013.07.012.

Zook, M. et al., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2), pp.6–32. Available at: http://doi.wiley.com/10.2202/1948-4682.1069.

Appendix

Appendix A: Results of Kriging interpolation methods per month

Appendix B: RMSE Boxplot tested interpolation methods October



Appendix C: Site rating of WOW-NL stations.

The WOW-NL stations have a quality rating which is based on various location attributes. The way the location attributes are used to determine this quality rating is based on the following classification system by the Met Office (Met Office 2016).

WOW lets you provide details about a number of attributes that help other WOW users and the Met Office understand the surrounding environment. These attributes have been compiled based on site grading schemes used by the Climatological Observers Link (COL), the World Meteorological Organisation (WMO) and the Met Office.

- **Exposure**
- **Measurements of air temperature**
- **Measurements of rainfall**
- **Measurements of wind**
- **Urban Climate Zone Index (UCZ)**
- **Reporting hours**

These Location Attributes are used to calculate a **Site's Rating**

Exposure

5: Very open exposure: no obstructions within 10h or more of temperature or rainfall instruments.

4: Open exposure: most obstructions/heated buildings 5h or from temperature or rainfall instruments, none within 2h.

3: Standard exposure: no significant obstructions or heated buildings within 2h of temperature or rainfall instruments.

2: Restricted exposure: most obstructions/heated buildings >2h from temperature or rainfall instruments, none within 1h.

1: Sheltered exposure: significant obstructions or heated buildings within 1h of temperature or rainfall instruments.

0: Very sheltered exposure: site obstructions or sensor exposure severely limit exposure to sunshine, wind, rainfall.

R: Rooftop site: Rooftop sites for temperature and rainfall sensors should be avoided where possible.

T: Traffic site: equipment sited adjacent to public highway.

U: Exposure unknown or not stated.

Exposure ratings relate to the site of the temperature and rainfall instruments only, which should ideally be at ground level. Sensors for sunshine, wind speed etc are best exposed as freely as possible, and rooftop or mast mountings are usually preferable.

Exposure guidelines are based on a multiple of the height h of the obstruction above the sensor height; the standard is a minimum distance of twice the height (2h). Thus for a raingauge at 30 cm above ground, a building 5 m high should be at least 9.4 m distant (5 m less 0.3 m, x 2), and a 10 m building should be at least 17 m from a thermometer screen (10 m less 1.5 m, x2).

Measurements of air temperature

A: Standard instruments in Stevenson Screen, calibration within last 10 yr, site exposure minimum rating = 3.

B: Standard instruments in Stevenson Screen or manufacturer supplied AWS radiation screen, calibration within last 10 yr, site exposure = 2 or 3.

C: Standard instruments in Stevenson Screen or manufacturer supplied AWS radiation screen, site exposure 1 or less.

D: Non-standard instruments and/or no or non-standard radiation screen and/or sheltered site, site exposure 1 or less.

U: Instruments unknown or not stated.

0: No air temperature measurements made at this site.

STANDARD INSTRUMENTS in this context means: Calibrated mercury-in-glass thermometers or calibrated electronic temperature sensors.

Measurements of rainfall

A: Standard ''five inch'' manually-read raingauge or calibrated tipping-bucket raingauge, at standard height above ground (30 cm), site exposure minimum = 3.

B: Standard ''five inch'' manually-read raingauge or calibrated tipping-bucket raingauge, the rim mounted at standard height above ground (30 cm), exposure = 2 or 3.

C: Standard ''five inch'' manually-read raingauge or calibrated tipping-bucket raingauge, the rim mounted at standard height above ground (30 cm), exposure 1 or less.

D: Non-standard raingauge and/or tipping-bucket raingauge, exposure 1 or less.

U: Instruments unknown or not stated.

0: No rainfall measurements made at this site.

STANDARD INSTRUMENTS in this context means: Standard-pattern (Snowdon or Met Office Mk II pattern) ''five-inch'' copper raingauge, with deep funnel, the rim of the gauge level and mounted at 30 cm above ground level, meeting the minimum exposure requirement of being at least 'twice the height'' of the obstacle away from the obstacle.

Measurements of wind

A: Wind sensors calibrated within last 10 years, mounted 10m above the ground on mast or pole, with no obstructions within 100m.

B: Wind sensors mounted above the ground on mast or pole, with no obstructions within 50m.

C: Wind sensors mounted on building or wall.

U: Instruments unknown or not stated.

0: No wind measurements made at this site.

Urban Climate Zone Index (UCZ)

1: Intensely developed urban zone with detached close-set high-rise buildings with cladding, e.g. downtown towers.

2: Intensely developed high density urban with 2 - 5 storey, attached or very close-set buildings often of brick or stone, e.g. old city core.

3: Highly developed, medium density urban with row or detached but close-set houses, stores & apartments e.g. urban housing

4: Highly developed, low density urban with large low buildings & paved parking, e.g. shopping mall, warehouses.

5: Medium development, low density suburban with 1 or 2 storey houses, e.g. suburban housing.

6: Mixed use with large buildings in open landscape, e.g. institutions such as a hospital, university, airport.

7: Semi-rural development with scattered houses in natural or agricultural area, e.g. farms, estates.

U: UCZ unknown or not stated.

*UCZ descriptions as defined by the **World Meteorological Organisation (WMO-No.8, 7th Edition)***

Reporting hours

A: Will always aim to provide a weather report at 09:00 GMT. Daily temperature and rainfall values relate to standard 24 hour period morning to morning.

B: Will always aim to provide a weather report between 06:00 and 09:00 GMT. Daily temperature and rainfall values relate to standard 24 hour period morning to morning.

C: Daily temperature and rainfall values relate to the 24 hour period midnight to midnight. This is the default for most automatic weather stations.

D: Air temperature and rainfall terminal hour is other than A, B or C above, or extremes do not relate to 24 hour periods.

U: Reporting hours unknown or not stated.

How site ratings are calculated

Each site is automatically allocated a 'site rating'' based on the observing location attributes entries submitted on site registration. The system is based on the quality and exposure of the temperature and rainfall data:

5* = E5, T=A, R=A
4* = E >= 3, T=A, R=A
3* = E >= 3, T[=A,B or C], R[=A,B or C]
2* = E >= 1, T[=Any], R[=Any]
1* = E =0,1,R or U, T[=Any], R[=Any]

(Where E = **Exposure**, T = **Temperature**, and R = **Rainfall**, and each of these are described in **Location Attributes**).

If temperature is measured at a site, but not rainfall, the site rating will be based on the quality and exposure of the temperature data alone. If rainfall is measured at a site, but not temperature, the site rating will be based on the quality and exposure of the rainfall data alone.

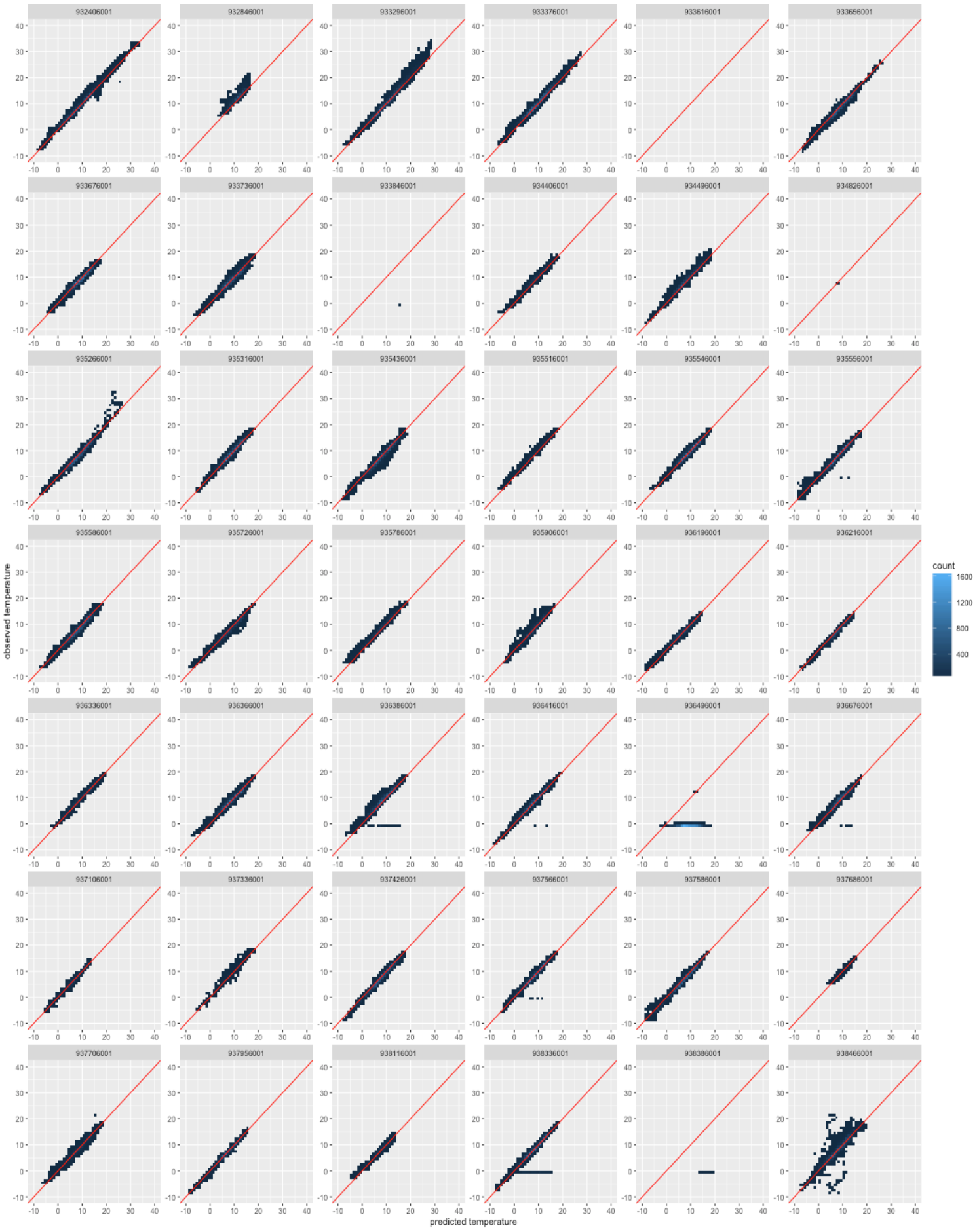
*If there is no temperature or rainfall data, the site will be classed as 1**

Appendix D: Individual station residuals (binned) for July, August, October, November, December, and January.


898396001
898436001
900306001
901646001
902806004
904886001
911206001
913926001
914096001
914176001
914196001
914256001
916136001
916156001
916206001
916236001
917816001
917826001
917896001
917906001
917936001
918216001
918396001
918426001
918716001
918736001
918746001
918756001
918766001
918776001
918796001
918826001
919666001
919726001
919736001
919756001
919836001
919846001
919866001
919926001
919966001
920116001
observed temperature
predicted temperature
count
1600
1200
800
400


920546001
920576001
920586001
920646001
920656001
920996002
921146001
921216001
921486001
921496001
921526001
921696001
921926001
922086001
922386001
922406001
922456001
922476001
922486001
922846001
923266001
923286001
923406001
923476001
923556001
923726001
923766001
923796001
923836001
923956001
923966001
923976002
924046001
924276001
924296001
924306001
924336001
924356001
924366001
924386001
924516001
924536001
observed temperature
predicted temperature
count
1600
1200
800
400


924956001
925376002
925486001
925516001
925986001
926186001
926356001
926656002
927206001
927226001
927246001
927356001
927906001
928016001
928306001
928556001
928566001
928826001
928846001
928876001
929126001
929236001
929366001
929856001
929876001
929886001
929966001
930006001
930146001
930196001
930206001
930246001
930296001
930936001
931326001
931726001
931746001
931786001
931826001
931946001
931996001
932256001
observed temperature
predicted temperature
count
1600
1200
800
400


932406001
932846001
933296001
933376001
933616001
933656001
933676001
933736001
933846001
934406001
934496001
934826001
935266001
935316001
935436001
935516001
935546001
935556001
935586001
935726001
935786001
935906001
936196001
936216001
936336001
936366001
936386001
936416001
936496001
936676001
937106001
937336001
937426001
937566001
937586001
937686001
937706001
937956001
938116001
938336001
938386001
938466001
observed temperature
predicted temperature
count
1600
1200
800
400


920546001
920576001
920586001
920646001
920656001
920996002
921146001
921216001
921486001
921496001
921526001
921696001
921926001
922086001
922386001
922406001
922456001
922476001
922486001
922846001
923266001
923286001
923406001
923476001
923556001
923726001
923766001
923796001
923836001
923956001
923966001
923976002
924046001
924276001
924296001
924306001
924336001
924356001
924366001
924386001
924516001
924536001
observed temperature
predicted temperature
count
1600
1200
800
400

Appendix E: Thin Plate Spline interpolation R-script written by Paul Hiemstra.

```
doTps = function(formula, data, newdata, ..., debug.level = 1, addFit = FALSE) {
        require(fields, quietly = TRUE)
        if(debug.level > 0) cat("[using thin plate splines (from fields)]\n")
        f = as.character(formula)
        dependent = f[2]
        independent = f[3]
        if(independent == "1") {
        fit = Tps(x = coordinates(data), Y = data[[dependent]], ...)
        newdata$var1.pred = as.numeric(predict(fit, coordinates(newdata)))
        } else {
        fit = Tps(x = coordinates(data), Y = data[[dependent]], Z = data[[independent]], ...)
        newdata$var1.pred = as.numeric(predict(fit, coordinates(newdata), Z =
        newdata[[independent]]))
        }
        newdata$var1.var = NA
        if(debug.level > 1) print(fit)
        if(addFit) {
        ret = list(krige_output = newdata, fit = fit)
        class(ret) = c("autoKrige", "list")
        } else {
        ret = newdata
        }
        return(ret)
}
```