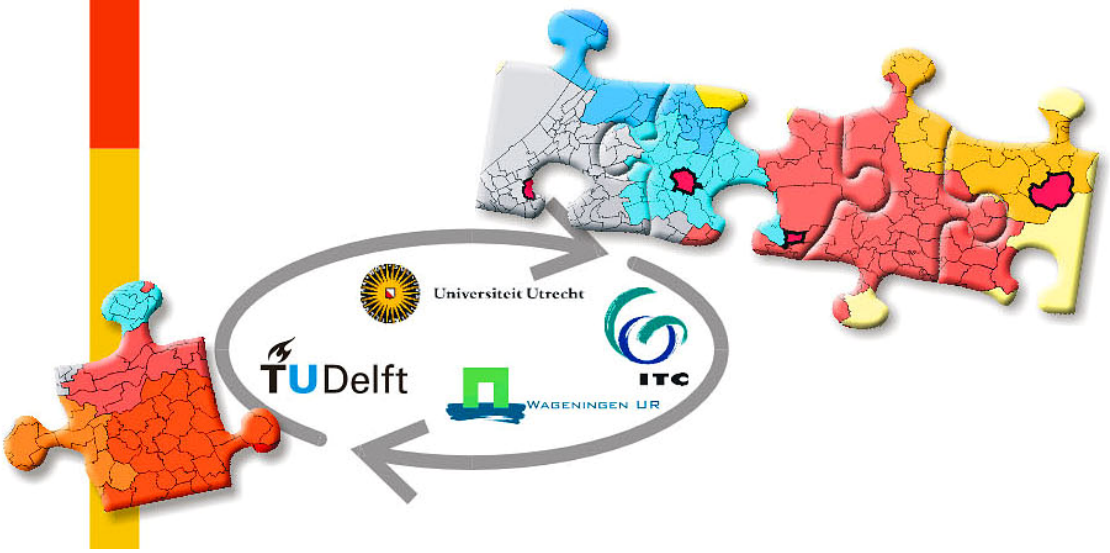
GIMA

Geographical Information Management and Applications

SPATIO-TEMPORAL DATA MINING FOR VINEYARD YIELD ESTIMATION

A Slovenian case study using linear regressions and self-organizing maps

MIRAN TISU



SPATIO-TEMPORAL DATA MINING FOR VINEYARD YIELD ESTIMATION

A Slovenian case study using linear regressions and self-organizing maps

MASTER THESIS

DATE: November 2012

AUTHOR: Miran Tisu

SUPERVISOR: Dr. Raúl Zurita-Milla (ITC Enschede)

REVIEWER: Dr. Ir. Sytze de Bruin (Wageningen University)

Abstract

Crop yield estimation is important for a number of reasons. For instance, crop yield estimation can be used to plan harvest and storage requirements. The main objective of this thesis was to use spatio-temporal data mining to estimate vineyard yield in Slovenia. For this, an after-harvest yield model capable of estimating and verifying yield at any location was built. Data mining was used to discover and quantify spatio-temporal relations between vineyard yield and selected explanatory variables. The Goriška Brda vine district (western part of Slovenia) was selected as study area and yield was estimated for all grape varieties in the district and for the Rebula variety, which is the most common variety in the district.

From a methodological point of view, the aim of this thesis was to explore different spatio-temporal data mining approaches. Thus, two types of regressions, namely the ordinary least squares (OLS) and the geographically weighted regression (GWR), and a type of neural network method, self-organizing maps (SOM), were explored.

The data available for this study mainly derives from the Slovenian Ministry of Agriculture and Environment (MAE). From the available data, explanatory and yield data was extracted. The explanatory variables were phisiogeographical (slope, exposition, etc.), vineyard characteristics related (distance between rows, distance between vines, etc.) and socio-economical (e.g. area of vineyards cultivated by a farmer). The dependent variable is the after-harvest grape yield per vine declared by the farmers. The spatial unit for this research is the single vineyard field and the time span of the data used in the research is five years, from 2007 to 2011.

OLS and GWR regression results were compared to identify the method that better explains grape yield variation. Regression results were also compared to selected meteorological characteristics to estimate their effect on the accuracy of yield estimation. After that an unsupervised SOM clustering was done (i.e. the dependent variable (yield) was not taken into account when performing the clustering). The resulting SOM clusters were projected into the geographical space in order to check for spatial patterns. Further, yield's variation within clusters was investigated to asses the value of the clustering.

Results for all three data mining methods and for all grape varieties as well as for the Rebula variety indicate that yield can not be properly estimated using the selected methods and / or the selected explanatory variables. When comparing the regression results, GWR reached better results than OLS although with an R^2 of approximately 0.15 – 0.25 (depending of the year). The comparison of prediction accuracy with meteorological characteristics shows no relation either. The SOM results were similarly poor. Clusters were barely identifiable and though they did form certain geographical patterns, they did not reflect yield variation (i.e. average yield is similar for all clusters). Possible causes for these poor results are a) the absence of a clear yield pattern in the dataset, b) the suitability of the explanatory variables, c) the data preparation and the parameterization of methods. Nevertheless the research provides the tools to identify declaration errors that should be inspected and corrected to improve MAE registers' accuracy.

Acknowledgements

First of all I would like to thank my supervisor, Raúl Zurita-Milla, for his valuable advice, wise direction and guidance, numerous corrections and availability at any time, though supervising me from a distance.

I would also like to thank Ministry of Agriculture and Environment of Slovenia (MAE) for datasets used in the research and to Mojca Jaška for explaining the registers.

Many thanks to my parents, Mirko Tisu and Fani Tisu, to my sister Mojca Tisu, to Vedran Jakačić and to Barbara Volkar for being there when I needed them the most.

Finally, immense gratitude to my partner, Polona Čelhar, for her limitless support and patience.
I would not have made it without you.

Table of contents

List of Figures

List of Tables

1. INTRODUCTION

The production of crop and yield estimation has direct impact on year-to-year national and international economies and plays an important role in food management (Prasad et al. 2006). In agriculture, yield estimation is important for example for planning harvest and storage requirements, for delivery estimates and for crop insurance purposes. In viticulture, yield estimation is important for the same reasons. One can however expose the importance to make plans for vintage and provide enough space for new wine (at farm level), and the importance for monitoring, to define production policies, taxes, etc (at the regional/national level). It is because of the latter, that it is important for a county or a region to also have accurate data about yield after it is harvested, especially if not all of it is declared.

Grape yield (yield further in the document) and wine production at the regional/national level has more impact than at the farm level because it is directly or indirectly relevant for more people (or companies). This is why EU member countries are obliged to report their yearly yield and wine production to The Commission of the European Communities. Reports at the national level are required during the growing season (estimation) and after the harvest (declaration) (European Commission, 2009).

In Slovenia, yield estimation/final declarations are based on field work (estimation during growing season) and on farmers' declarations (final declaration after harvest). Yield estimation methods using crop models, which are common for estimating yield of other crops (e.g. maize, soy, cotton) in agriculture during growing season, have not been considered for vines yet. Alternative methods for estimating (non-declared) yield after harvest have also not been considered yet. The latter is mainly because estimation of overall yield (declared and undeclared one) can be derived from farmers' declarations. Its accuracy, however, depends heavily on the accuracy of the farmers' declarations, on the ratio of declared yield and on the assumptions of undeclared yield made by the Ministry of Agriculture and Environment of Slovenia (MAE). Nevertheless, alternative yield estimation methods should be properly tested before they are applied, as it is not certain that they would actually improve the current level of accuracy.

For the purpose of improving the accuracy of yield estimation, it could prove useful to construct a yield estimation model capable of estimating/verifying yield at any location, regardless if it has been declared at that location or not. This way, one could verify the yield declared by the farmers and estimate the yield of undeclared locations. Yield at national level could be estimated by adding up these estimations. If such a model explained high levels of variation of the declared yield, a step towards estimation of yield during growing season could be researched next. Similarly to other crop yield estimation methods, sensor data, national databases and other kinds of available ancillary geo-datasets are expected to provide valuable information to construct such models.

1.1 Background

Out of 21 500 ha of vineyards cultivated in Slovenia in 2011 (land use ortophoto interpretation), 16 000 ha were declared in Slovenian register of grape and wine growers by farmers. Yield declarations however only cover about 13 000 ha, just over 60 % of the area of all vineyards. Furthermore, the distribution of declared vineyards and declared yield is not even across the country. The ratio of declared yield is generally lower in areas where the average size of vineyards is small and/or in areas where overall vineyard area per farm is small (Jakša, 2011). The request of the European Commission for mid-season (estimation) and after-harvest (declaration) report of

yield (European Commission, 2009) therefore presents a challenge for Slovenia because of the relatively high ratio of missing data and its irregular spatial distribution.

For mid-season yield estimation, data sampling techniques are used in Slovenia. Data is sampled by specialists in selected vineyards from all wine-growing regions. Rules on methodology and location of sampled sites are defined in the national legislation (Uradni list, 1999a). The results of the sampling are used to estimate yield at wine region level. Estimations are reported to the MAE, which adds them up to estimate yield at national level (Jakša, 2011).

The amount of harvested yield is partly estimated, as all yield is not declared by farmers. Therefore the amount of overall yield (declared and undeclared) is in this document referred to as after-harvest yield estimation. That is in fact the yield quantity that is monitored for national policy purposes and is reported to the European Commission as yield declaration at national level. It is however based on the yield data submitted by farmers. The overall yield is estimated to be 143 % of the declared yield to account for self consumption, which according to the estimations made by the MAE amounts to approximately 30 % of all yield. A relatively low increase factor (of 30 %) is used compared to the area of vineyards without yield declaration (almost 40 % of area of estimated vineyards). This is because undeclared yield is mostly produced in small, self consuming and, in many cases, old vineyards with fewer plants per area unit, which consequently produce less yield. This calculation is made at the national level, not taking into account the geographic location of the vineyards with missing declarations (Jakša, 2011).

Current after-harvest yield estimation methods in Slovenia do not take into account, or if they do it is only partly, a number of factors which could, according to literature (Vršič & Lešnik, 2001; Stevenson, 2005; Gouveia et al., 2011; Rusjan & Korošec-Koruza, 2003) influence the quantity (and quality) of yield. By estimating yield at the level of vineyards (to be added up to derive yield at the regional/national level), such factors can be taken into account. The most important of such unaccounted factors are:

- **Natural conditions (exposition, climatic conditions, etc.),**
- **Properties of vineyards (distance between plants, age of vines, etc.),**
- **Location of vineyards (x, y, z).**

Natural conditions can be divided into (relatively) permanent (exposition, slope, soil, etc.) and into those that vary each season (precipitation, temperatures, etc.). The permanent natural conditions change intensively and non-uniformly form location to location in the hilly type of relief, which is the most common for vineyard plantations in Slovenia. The varying natural conditions are also related to the location but they do not change as intensively as the relief conditions. On the other side, they differ every year and, according to literature (Shanmuganathan et al., 2010) this makes them crucial factors for yield quantity and quality.

The **properties of a vineyard**, according to Vršič & Lešnik (2001), do influence yield, as they determine, for example, the area available for a certain plant (vine and row distance) or the expected yield ratios (low yield in first few years). Furthermore, there is some variation between the expected yield of different grape varieties. Thus, by taking into account vineyards characteristics, one should be able to more accurately predict their yield.

The information about **the location of a vineyard**, could improve estimation accuracy as well. In this case, the influence of the vineyard's micro location would be considered. The natural characteristics of a vineyard, such as exposition, altitude, climatic conditions etc. often derive from, or are dependant on, the vineyard's location. The vineyards' micro location though could bare additional information that is not included within the variables mentioned above, for example

micro-climatic properties that are not distinguishable from coarse meteorological data or from relief properties.

Raw data (the data as acquired), might not be the only source of explanatory variables. Preprocessing the data to obtain additional variables might prove useful as well. For instance, if the spatial unit is a single vineyard with its characteristics, the information about quantity of all of the farm's vines or the information about the area of all of the farm's vineyards might be useful as well. Such socio-economic characteristics of a farmer / farm might influence the goal and the experience of the farmer and thus the yield per plant on that particular farm.

However, the relations between input variables and yield have to be identified and quantified in order to construct a yield estimation model. Existing crop yield estimation models rely on various modelling approaches and input data. Here, only a few are mentioned in order to provide a brief overview (more details are provided in chapters 2 and 4). For instance, Prasad et al. (2006) constructed an overall yield estimation model for corn and soy bean for Iowa (US) based on normalized difference vegetation index (NDVI), soil moisture, surface temperature and rainfall data. Everingham et al. (2009) constructed a relatively complex model based on climatic data to estimate sugar cane crop yield in Australia months before its harvest. Gouveia et al. (2011) constructed a model for predicting wine quantity in the Douro valley (Portugal). This model can be used in the early and mid growing seasons and uses monthly means of climate variables and of NDVI as an input. Despite their value, all these models only focus on the overall yield of a region. Cao et al. (2011) constructed a model based on soil nutrient content that was used to predict maize yield per grid cells in fields of approximately 375 acres in China.

Vineyard yield estimation accuracy in Slovenia might benefit from using such crop modelling approaches. There are several datasets available with a potentially high explanatory power. However, relevant information still has to be extracted from the data. A possible way to extract such information is data mining, the search for hidden patterns in large databases (Ng & Han, 2002). Data mining is actually behind the yield estimation models mentioned above. It was used in order to identify the explanatory variables that have important relation to the dependent variable, and to quantify their relations.

1.2 Research objectives

The main objective of this thesis is to use spatio-temporal data mining to estimate vineyard yield. Our goal is to build an after-harvest yield estimation model capable of estimating/verifying yield at any location, regardless if it has been declared at that location. In particular, data mining is used to discover and quantify spatio-temporal relations between vineyard yield and selected explanatory variables. The relations between the yield of Rebula grapes (the most common grape variety in the study area) and the available explanatory variables are studied too. For this study, yield in kg per vine is regarded as the dependant variable, single vineyard fields are used as the spatial unit of analysis and the data from 2007 to 2011 from the Goriška Brda Slovenian wine district are available.

From a methodological point of view, the aim is to explore different data mining approaches. Thus, two types of regressions, namely the ordinary least squares (OLS) and the geographically weighted regression (GWR), and a type of neural network method, self-organizing maps (SOM), are explored. Both OLS and GWR are linear regression methods. However, OLS is a global regression method while GWR is a local regression method (i.e. it takes location into account when estimating the

dependent variable). SOM is a type of unsupervised artificial neural network clustering method and, as such, this approach is completely different than OLS and GWR. Finally, a temporal component is added to the data mining problem by applying the methods to the time span of five years and by comparing the accuracy of the results of various years.

1.3 Research questions

The above mentioned research objectives lead to the three main research questions of this thesis:

Can spatio-temporal data mining as committed within this research:

- **Provide the information suitable to construct a model for after-harvest yield estimation in Slovenia?**
- **Provide the most suitable method for constructing such a model?**
- **Provide the basis for mid-season yield estimation in Slovenia?**

Here, yield means both overall yield as well as the yield of the most common grape variety of the area of research (Rebula) and, as stated in section 1.2, yield estimation is done at the scale of individual vineyard fields. In order to answer these main research questions, research sub-questions will have to be answered:

- *Which available variables affect the quantity of yield most?*

In order to estimate yield, the variables that significantly explain the variation of yield need to be identified. Environmental variables, such as slope, exposition, etc. and vineyard characteristics, such as age of vines, distance between plants, etc. are considered. Socio-economic variables, such as number of vine plants per farm and area of vineyards per farm, which are derived from raw data, are considered as well.

- *Which of the two regression methods, OLS or GWR, estimates vineyard yield better?*

OLS is a common method to predict yield and GWR is a relatively newer method with an ability to include geographical components in the regression model and, consequently, to asses the influence of location on the regression's accuracy.

- *Can regression methods accurately estimate vineyard yield?*

It is important whether the level of accuracy of the most accurate method is high enough to asses a method as feasible to estimate yield. It is also important whether there is any difference in prediction power for entire yield versus the prediction power for one grape variety.

- *Can the SOM method successfully cluster available explanatory data?*

The answer to this question is important to asses the suitability of the SOM method for our problem. Though SOM is in its core a clustering method, a variation of SOM, the supervised SOM, has a prediction capability, and could be applied to estimate yield in case that the SOM method proves feasible.

- *Is yield quantity reflected in clusters derived from SOM and do these clusters have a geographical pattern?*

SOM clusters should be related to yield in order to be useful for our purpose. Further, a geographical distribution of clustered data records should be analyzed in order to observe whether there is a geographical pattern in the spatial distribution of the different clusters.

- *What is the relationship between yield estimation accuracy and meteorological characteristics?*

A temporal period of five years is considered in this thesis. The question is whether the meteorological characteristics of particular years do influence the accuracy of yield estimation. If this is the case, this is important for future mid-season yield estimation studies.

- *Can new findings be applied to research mid-season yield estimation?*

An estimation of applicability of mid-season yield estimation, using the same data and methods, is possible after the results of after-harvest yield estimation are analyzed.

2. LITERATURE REVIEW

In this chapter, the most important facts and methods regarding this research are reviewed. The chapter begins with a brief description of viticulture and yield estimation from the perspective of this research. The focus of these topics is on description of viticultural facts and on yield estimation techniques that are important for constructing the methodology of data mining approach in our research. Data mining and knowledge discovery in general is presented next. Finally, the description of data mining methods that are applied in this research is given. The description of ordinary least squares (OLS) is followed by the description of geographically weighted average (GWR) and self-organizing maps (SOMs). The description of the data mining methods includes the description of the methods, the means to asses its results and, with exemption of OLS, the examples of their application as found in literature.

2.1 Viticulture

Common grape vine (Vitis Vinifera) ('vine' further in the document), is a domesticated plant, grown in plantations (vineyards) mostly for the purpose of making wine. The science or practice of growing grapes, especially for the purpose of making wine, is viticulture (Hrček & Korošec-Koruza, 1996).

Figure 1. A row of vines in a vineyard in Goriška Brda.

Vine is a perennial plant that can live over 400 years; the oldest known living plant is located in Maribor in Slovenia (Records G.W., 2011). However, vines in plantations, grown for the purpose of wine making, are usually up to 30 years old and get replaced by new ones as their yield ratios decrease. In the first few years after planting new vines, the primary goal is to grow a bearing and

healthy vine. This requires special and more intensive work with the plants. Also, there is no yield in the first year while in the next few years the yield is low (Vršič & Lešnik, 2001).

Natural factors that influence the growth of vine are climate, soil, location and topography. Suitable annual mean temperatures for growing vine are between 10 and 20 degrees Celsius. The distribution of mean and extreme temperatures is important, as vine is sensitive for spring frost and produces better and more abundant yield if summer is warm. Vine needs approximately 1300 – 1500 hours of sunshine and around 700 mm of rainfall during the growing season. Ideally rain periods would be during winter and spring. Furthermore, the selection of the right soil is important because it influences drainage levels and the amount of minerals and nutrients that vine is exposed to. Ideally, the soil would retain water sufficiently, but would also have good drainage, so the roots would not become overly saturated (Stevenson, 2005). When choosing a proper location and topography for a vineyard, one has to consider altitude, slope, exposition, closeness to nearby lakes and rivers, latitude, etc. These factors influence micro-climatic characteristics of an area and are thus particularly important. They influence the amount and strength of sunlight received by plants as well as temperature, precipitation and wind characteristics of location. They also determine the tendency for occurrence of diseases and pests and finally the cost of production (Vršič & Lešnik, 2001).

There are a number other factors that have to be considered by a farmer when planting a vineyard. The natural factors mentioned above have, for example, different impact on different grape varieties, as their need for optimal conditions differ. Ideally, particular grape varieties should be planted in the area, where ecological characteristics would enable selected grape varieties to fully exploit its genetic potential to provide its highest quality grapes (Rusjan & Korošec-Koruza, 2003). Besides optimal conditions, resistance from weather extremes, diseases, etc. are different by varieties as well. Further, vines in vineyard are trained in order to maximize the amount of sunlight received by the plant. They can be trained in a number of training systems that allow vintners to manage the canopy in an optimal way (according to slope, variety, etc) and to control the yield of vine. Figure *2* shows two possibilities of the same training system. Finally, vines are grafted onto different rootstocks and the type of rootstock used for grafting has an influence on a number of vines characteristics, for example on resistance to drought, on timing of grape ripening, etc. There are a number of possible combinations of variety, training system and rootstock. The farmer's task is to choose the right combination according to vineyards natural characteristics. Besides that they must provide balanced fertilizing, protection from diseases and pests, etc. (Vršič & Lešnik, 2001, Hrček & Korošec-Koruza, 1996).

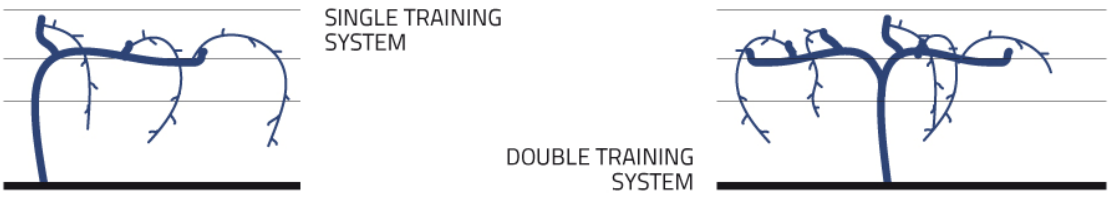


Figure 2. Single and double training systems of the same type.

Properties of grapes are crucial for ingredients that determine the quality of the vintage, and are responsible for wine colour, aroma and flavour phenols. Vintage quality is very important for the winemaker, who eventually uses grapes to produce vine. Another very significant influence on the quality of wine is the winemaker's experience and talent (Shanmuganathan et al., 2010).

Because the properties of grapes also depend on the quantity of grapes per vine, some producers decrease quantity of yield in potentially more abundant years by decreasing the number of

clusters on a vine before the grapes mature (Hellman, 2003). A different way of yield decrease, applied because it is supported by EC policy, is removal of all grapes from the vineyard before they mature (green harvest). In EU member states who implemented this measure, farmers that meet certain conditions can apply for subventions in case they decide to perform green harvest on (parts of) their vineyards (Jakša, 2011).

Good results in yield quantity and quality require continuous work in the vineyard, even in the vine's dormant period. The duration of annual growth cycle of vine, from bud break to physiological maturity of crops, lasts from 95 to 120 days, even up to 170 days for late varieties. The beginning and ending of certain phases in the growth cycle, as well as their duration, depends mostly on the climate and on grape variety characteristics. Harvest, the removal of the grapes from vine, is usually performed in time of physiological maturity of crops, which usually occurs in autumn. Its timing is mostly dependent on climate conditions and grape variety characteristics. The growing phase of vine is followed by the winter dormancy period, which starts when leaves begin to fall and the vine stops growing (Vršič & Lešnik, 2001).

2.2 Yield estimation

Yield estimation is predicting the quantity of yield. An experienced farmer can predict the yield of his crops relatively accurately, but the accuracy of estimation can become difficult for a farmer or a company if a lot of land is being farmed. It is also difficult to estimate yield at regional or national level, as there are often many variables and local/regional characteristics to be considered. With rapid development of technology in last decades, more accurate yield estimation at regional and national levels has become possible. Particularly important for this progress are new possibilities of data acquisition, namely automated data acquisition. Automated data acquisition can in general be divided into data acquisition when a sensor makes a physical contact with a measured phenomenon, and into remote sensing data acquisition, when data is acquired without such contact. As different kinds of data became available, new approaches and techniques for yield estimation have occurred.

Some researchers estimate yield using data deriving from sensors stationed on the earth's surface. For example Shanmuganathan et al. (2010) modelled vine crop quantity and quality of a vineyard in New Zealand using various meteorological data from automated sensors of nearby meteorological station. Unganai & Kogan (1998) on the other hand managed to estimate corn yield and monitor drought using data obtained from the Advanced Very High Resolution Radiometer (AVHRR) sensor on board the satellites. The research of course included validation using precipitation and yield data gathered on the earth's surface. Finally, some researchers, that were already mentioned in chapter 1 (Gouveia et al., 2011; Prasad et al., 2006), used data deriving form both direct and remote sensors in order to estimate yield.

Yield is usually estimated at regional level. Regional crop forecasting procedures can be categorized into two strategies, "bottom-up" and "top-down". Bottom-up approaches consider components that influence the yield at very detailed level of the system and gradually merge them to predict yields on a larger spatial domain. This kind of approach on the one hand provides verification and estimation at parcel level and on the other hand the estimation at regional level. The disadvantage associated with the bottom-up approach is that the errors from smaller scales can aggregate. The alternative top-down approach on the other hand considers the major system components that contribute to yield and integrates the information at more detailed levels if

required. Its disadvantage is the possibility of oversimplification by assuming regional homogeneous environmental and management conditions (Everingham et al., 2009).

Yield can be estimated during the growing season, when its quantity is not yet known, or after the growing season, if all of its quantity has not been measured or declared. Estimation of yield during the growing season, or mid-season yield estimation, is common at farm level as well as at regional/national level. Estimation of yield after harvest (if whole amount of yield is not known), or after-harvest estimation, is common at regional and national level, because it is a source of information for monitoring purposes and for policy making (Gouveia et al., 2011).

After-harvest grapevine yield estimation is not often reported in literature. An example of after-harvest grapevine yield estimation is research that was performed by Shanmuganathan et al. (2010), who applied data mining in order to classify a single vineyard's yield to three ordinal classes (low, medium, high). The estimation was based on meteorological data that was gathered over a period of 12 years.

The research and modelling of mid-season yield estimation is most common. Under the description below, examples of mid-season vineyard yield estimation at farm level are followed by the examples of mid-season yield estimation of wine and some common agricultural crops at regional level. Examples present estimations for large regions as well as for areas of smaller scales. Some examples were already briefly mentioned in section 1.1.

Mid-season yield estimation at vineyard or farm level is common in viticulture. It is important for viticulturists and vintners in order to make necessary plans for a vintage (Wolpert & Vilas, 1992) and/or for achieving target yields (Hellman, 2003). Basic technique for mid-season yield estimation at the level of vineyard is based on the number of vines, data sampling and historical yield data. More advanced technique is based on data sampling in certain growth phases and uses an "increase factor" in calculation instead of historical yield data (Hellman, 2003). However, techniques which rely on automated data acquisition have been introduced lately. Blom P.E. & Tarara J.M. (2009) propose trellis tension monitoring technique, where deployed systems provide dynamic measurement of changes in the tension of main trellis support wire in a vineyard. Another automated technique is based on automated sensor-based approach for sampling, in this case detection and counting of grapes. Images are collected from a camera mounted on a vehicle, which is driven along the rows in a vineyard. Algorithms are used to predict yield from sensor-based sampled data (Nuske, 2011).

It is more common in agriculture to estimate yield in mid-season at regional or national level as yield forecasts are pivotal for the success of any agricultural industry that plans or sells ahead of the annual harvest. A diverse range of agricultural industries, such as wheat, corn, maize, cotton and others rely on accurate and timely crop forecasts (Everingham et al., 2009). Such forecasts or estimations, as well as estimations at lower levels, often deploy data mining methods in order to provide results.

Gouveia et al. (2011) managed to construct robust and reliable linear regression models for wine estimation in the area of Portuguese Douro Valley. Two models were constructed, based on selected monthly means of climate variables and normalized difference vegetation index (NDVI). The first model predicts yield in early season (March) and the second in mid-season (July). They both perform well, the mid-season model for example explains 90 % of variance. However, the main drive to construct these regional level models was not to estimate yearly yield, but to use it in order to estimate yield in mid-future. The goal was to compare climate scenarios predicting different growing conditions as a consequence of climate change. Everingham et al. (2009) used ensemble data mining approaches to forecast regional sugar cane production in Australia.

Ensemble data mining is machine learning method that leverage the power of multiple models to achieve better prediction accuracy than any of the individual models could on their own. The resulting model produced a high predictive correlation (Rcv) of 0,71 when predicting end of season sugarcane yields approximately 4 months before the start of the harvest and 10 months before harvest completion.

Ferraro et al. (2009) on the other hand performed an analysis of factors that influence sugar cane yields on the example of six farms located in northern Argentina. They used data mining techniques, such as classification and regression trees in order to derive the main factors that influence sugarcane yield per hectare. Data over the period of 5-years were used to obtain the relations between yield rations and weather conditions. Cao et al. (2011) on the other hand applied spatio-temporal data mining techniques in order to predict the yield of maize in the Yushu Gongpeng village on the level of a 40x40 m grid. The overall area of research covered approximately 375 acres while the data used covers a period of six years. They applied the neural network method to calculate the influence of soil nutrient content on yield. Furthermore, they constructed a linear regression model to obtain the integrated forecast outcome. The predicted values of the regression model differed from the actual values for less than 5 % for every year.

2.3 Data mining and knowledge discovery

"Data mining in general is the search for hidden patterns that may exist in large databases" (Ng and Han, 2002, p. 1003). The notion of finding useful patterns in data has been given a variety of names, but "data mining" as a phrase most frequently used by statisticians, data analysts and management information systems communities, was eventually adopted (Fayyad et al., 1996).

The traditional method of finding patterns in data and turning data into knowledge is based on manual analysis and interpretation of results. When data were not as abundant as today, these methods were in most cases sufficient. In past decades however, the volume of data has increased enormously. Such increase of data volumes has become possible with the development of a number of fields of science, but above all because of the rapid development of computer technology. This provided capabilities for storage of enormous volumes of data, as well as the computational power to manipulate it. The traditional methods of manual analysis can not cope with amounts of data available today and are therefore unpractical. They are slow, expensive and highly subjective comparing to computational methods (Fayyad et al., 1996).

The quantity of agricultural data or agriculture related data has followed the trends mentioned above. In viticulture for example, vineyard information such as slope, exposition, altitude etc. can be derived if the vineyard location and a digital terrain model are available. Furthermore, large amounts of vineyard properties data (a register of such data is obligatory for example for EU member states) exist. Certain interesting or useful correlation with for example meteorological data or even demographical data might be discovered if analyzed. But such data are too voluminous and too complex to be analyzed using traditional methods.

Statistical tools and methods were traditionally used to obtain useful patterns form data, therefore the statisticians' view on this matter needs to be introduced as well. According to Hand (1998), statisticians have adopted methods for data manipulation and statistical analysis to computer-based data analysis techniques in 1960s. Computers have since then been used to fasten and ease statistical computations. These methods however have at their core remained the same. Statisticians use validated data, meaning that it meets certain quality demands and follows

certain rules that need to be applied to statistical analysis. Data or even database can be created mainly for the purpose of such analysis. With the increase in size and complexity and with the possibility of interlinking of databases, databases, which are primarily created for various other reasons, are now seen as a resource as well.

Statisticians are aware that there is much valuable information in databases and they acknowledge data mining as the set of tools by which this information may be extracted. They define data mining as a process of analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners. They however stress that the data in such vast databases might have flaws from the point of view of statistical analysis. Some of their concerns are (Hand, 1998):

- Contaminated data (outliers, missing data, checking of questionable data at its source),
- Nonstationarity (population drift – changing of underlying population, gradual distortion of measurements, ...),
- Selection bias (distortion of selected or available sample away from a simple random sample).

Statisticians see the promise and the opportunities of data mining as obvious. On the other hand, they are concerned with the fact that such vast amounts of data, as are used for data mining, can not be acquired and prepared for the analysis in the manner usually used for statistical analysis. It is practically certain that such data are invalid in some way. They also warn that it is necessary to consider whether the patterns discovered from the data are real (rather than chance fluctuations in the database), how to make valid probability statements about it (given the probably non-random nature of the data), and whether it is nontrivial, interesting and valuable (Hand, 1998). Fayyad et al. (1996) also warns of blind application of data mining methods (criticized as 'data dredging" in statistical literature), as a dangerous activity, easily leading to the discovery of meaningless and invalid patterns. Data mining carried out without regard for statistical aspects of the problem should therefore be avoided.

The question of when data mining is successful is analyzed by Malone et al. (2005). They define successful data mining as "The process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases" (p. 12). They further clarify this definition by explaining the characteristics of successful data mining:

- Non triviality – complex processing (rather than simple computations) is required to uncover the patterns that are buried in the data,
- Validity - the discovered patterns should hold true for new data,
- Novelty – the discovered patterns should be new,
- Usefulness – the discovered patterns should be useful for researchers or organizations,
- Comprehensiveness – the discovered patterns should be understandable to users and to add to their knowledge.

A crucial factor for success of data mining is the choice of appropriate data mining methods and algorithms. Fayyad et al. (1996) points out that a large portion of data mining application efforts can go to properly formulating the problem (asking the right question) rather than into optimizing the algorithmic details of a particular data mining method. There is no universal data mining method, thus choosing an appropriate method and algorithm for certain applications requires some insight into most suitable methods for a particular domain.

According to Fayyad et al. (1996), data mining, though being an important part of the process, refers only to a particular step in the process of discovering useful knowledge from data. They

name this process as knowledge discovery in databases (KDD). They identify steps prior and after data mining as crucial in finding a meaningful result. The basic steps in KDD, some of which are presented in Figure 3, are the following:

- Developing and understanding the application domain and identifying the goal,
- Creating a target data set by selecting a data set or variables on which discovery will be performed,
- Data cleaning and preprocessing, including removing noise if appropriate and deciding on strategies for handling missing data fields,
- Data reduction and projection, resulting in finding useful features to represent the data,
- Matching the goals from 1st step to a particular data mining method,
- Choosing the data mining algorithm(s) and selecting method(s) for searching patterns in data,
- Data mining,
- Interpreting mined patterns and possibly returning to steps 1-7 for further iteration,
- Using the knowledge directly, incorporating it into another system or documenting it.

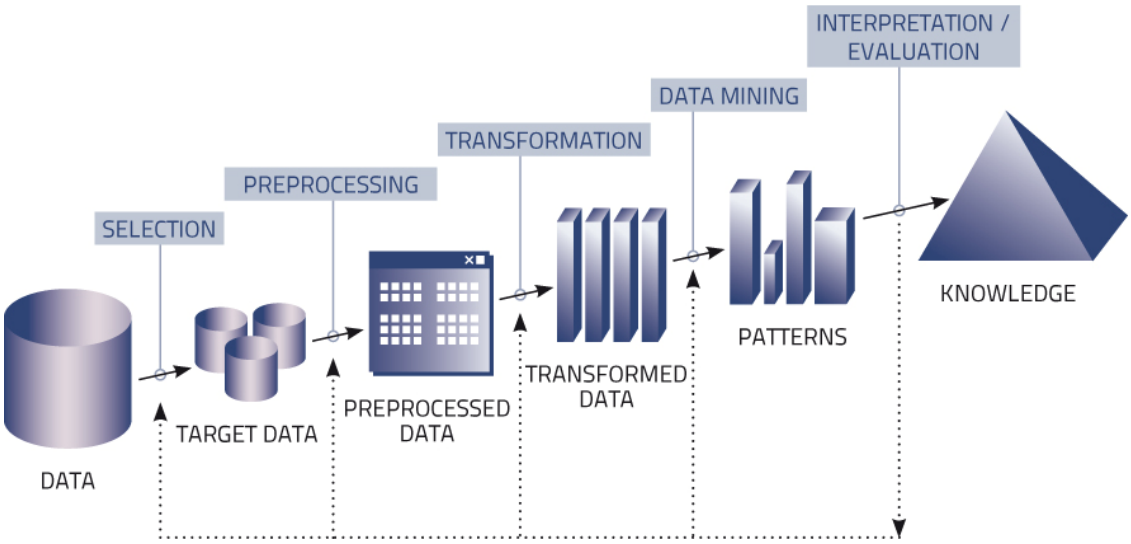


Figure 3. Steps in KDD process. Source: Fayyad et al., 1996, p. 41

According to Menins & Guo (2009), data mining and knowledge discovery is exploratory in its nature and is more inductive than traditional statistical methods. It fits in the initial stage of a deductive discovery process, where researchers develop and modify theories based on the discovered information from observation data.

Data mining and knowledge discovery are potentially useful in a number of scientific and commercial fields. In 1996, Fayyad et al. identified astronomy as one of the primary application areas. They also describe the usage of these methods in business, for example in marketing, investment, fraud detection, manufacturing, telecommunications and data cleaning. Over the years, data mining became an established method in scientific research. One can make an estimation of its use in various scientific fields by considering the number of published scientific articles by field. When we conducted an experiment by searching "data mining" in the well known academic citations index 'Web of knowledge" (WoK, 2012) in February 2012, computer technologies, genetics and medicine were the fields of science with the highest number of hits. According to article titles, data mining methods are very commonly used in DNA, genome and disease related research.

As with other kinds of data, the volume of spatial data is nowadays increasing enormously, as much more diverse, dynamic and detailed data can now be acquired using modern data acquisition methods. These methods are global positioning systems (GPS), high-resolution remote sensing, location aware services and surveys, internet based volunteered geographic information, etc. Besides geography related scientific fields, private industries and the general public have enormous interest in both contributing geographic data and using the data resources for various application needs (Menins & Guo, 2009). Similarly, the volume of temporal data has increased enormously using modern data acquisition techniques. Nowadays, temporal data can be obtained by registering events (telecommunication or web traffic data), monitoring processes and workflows, etc. (Andrienko et al., 2006).

Data mining using spatial and/or temporal data is referred to as spatial / temporal / spatio-temporal data mining (Andrienko et al., 2006; Menins & Guo, 2009). The need to investigate spatial and temporal relations at the same time complicates data mining tasks. Firstly, spatial relations (distance, topology, shape, etc.) and temporal relations (before and after, etc.) are information bearing and need to be considered in the mining techniques. Secondly, some of these relations are not explicitly coded in the database and therefore need to be extracted from the data. Thirdly, as working with the level of stored data is often undesirable, complex transformations are required. Fourthly, spatial resolution and temporal granularity have to be taken into consideration, as they can have a direct impact on the strength of patterns discovered. Finally, the approach in data mining differs on the domain of knowledge (Andrienko et al., 2006; Yao, 2003).

2.4 Ordinary least squares (OLS)

According to Hutcheson (2011), the ordinary least-squares (OLS) regression is a generalized linear modelling method that can be used to model a dependent variable. This method is one of the major techniques to analyze data and forms the basis of many other techniques (for example ANOVA and the generalized linear models). As such, OLS regression is used in a wide variety of fields.

The OLS method can be applied to single or multiple explanatory variables and also to categorical explanatory variables that have been appropriately coded. At a very basic level, the relationship between a response variable (y) and an explanatory variable (x) may be represented using a line of best-fit, where y is predicted, at least to some extent, by x. If this relationship is linear (as shown on Figure 4 – left panel) it may be appropriately represented mathematically using the straight line equation:

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

where β_0 indicates the value of y_i when x_i is equal to zero (also known as the intercept), β indicates the slope of the best-fit line (also known as the regression coefficient) and ε_i is the residual. The regression coefficient β describes the change in y that is associated with a unit change in x.

The deviation or difference between the observed and the predicted y values (residual) provides an indication of how well the model predicts each data point. Adding up the deviations for all the data points after they have been squared (to remove negative deviations) provides a simple measure of the degree to which the data deviates from the model overall (Figure 4 - right panel).

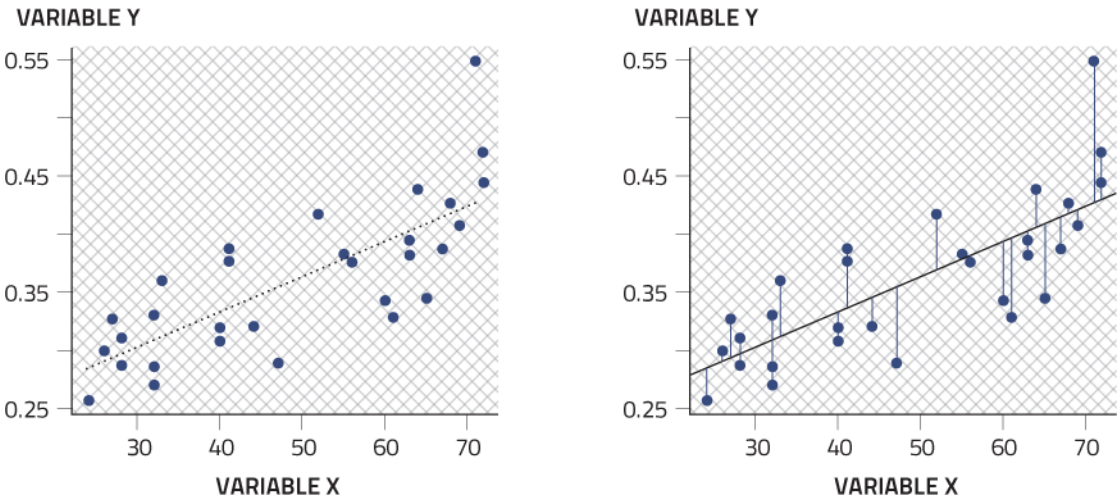


Figure 4. OLS regression line (left) and OLS regression model residuals (right).

The sum of all the squared residuals is known as the residual sum of squares (RSS) and provides a measure of model-fit for an OLS regression model. A poorly fitting model will deviate much from the data and will consequently have a relatively large RSS, whereas a good-fitting model will not deviate much from the data and will consequently have a relatively small RSS (close to zero).

However, the R^2 statistic provides more insight into (the performance of) the model. R^2 is also known as the coefficient of determination and it indicates the percentage of variation in the response variable that is explained by the model. It basically provides a measure of how well future outcomes are likely to be predicted by the model. Values of R^2 are between 0 and 1. Value 1 means that the regression line perfectly fits the data, while values close to zero mean that the model can not explain the variation of the dependent variable. R^2 is defined as:

$$R^2 = \frac{RSS \quad after \ \ regression}{total \quad RSS}$$

The OLS regression model can be extended to include multiple explanatory variables by adding additional variables to the equation. The form of a model is the same as above, but here a response variable (y) is predicted by multiple explanatory variables (x_1 to x_3):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

The interpretation of the parameters is basically the same as in a model with a single explanatory variable, but in this case the relationship cannot be graphed on a single scatter plot. β_0 indicates the value of y when x_1, x_2 and x_3 are zero. Each β parameter indicates the average change in y that is associated with a unit change in x, whilst controlling for the other explanatory variables in the model. Model-fit can be assessed through comparing deviance measures of nested models.

The R^2 statistic is used to derive the ratio of variance of dependent variable which is explained with a model. The adjusted R^2 (R^2 Adj) statistic can be applied to assess a model with multiple explanatory variables. It is a modification of R^2, dependent on the complexity of a model. A model with the same R^2 and less explanatory variables would result in a higher R^2 Adj. R^2 Adj is always lower than R^2.

A regression model should include explanatory variables that explain as much as possible the variance of the dependent variable. However, using too many explanatory variables might result in undesired effects. Particularly, when a large number of independent variables are incorporated in

a regression model, multicollinearity might be a problem. Multicollinearity is a high correlation of explanatory variables in a regression model. It occurs for example if the explanatory variables measure the same concepts or phenomena.

If the explanatory variables are collinear it is hard to measure the effect of a single variable on the dependent variable. One of the measures of multicollinearity is variance inflation factor (VIF). VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity of variables in the model. VIF is calculated:

$$VIF = \frac{1}{1 - R_i^2}$$

where R^2i is the coefficient of determination of a regression equation where explanatory variable (xi) is regarded as a dependent variable, explained by all of the remaining explanatory variables of original OLS model. The magnitude of collinearity is analyzed by considering the value of the VIF.

2.5 Geographically weighted regression (GWR)

As mentioned in the previous section, the focus of regression analysis is to find a relationship between a dependent variable and one or more independent variables. The output of a regression analysis is a regression function that describes such relation. In spatial analysis, however, the data are drawn from geographical units and if applying standard (i.e. OLS) regression, a single regression equation is derived. This produces global parameter estimates, which are assumed to apply equally over the whole region. Therefore such approach is referred to as global regression. The relationships being measured using this approach are assumed to be stationary over space.

According to Fotheringham et al. (1998) it is reasonable to assume that relationships can vary over space and that the parameter estimates might in some cases exhibit significant spatial variation. He acknowledges three main reasons for such variation. First, the parameter estimates can vary because of random sampling variations in the data used to calibrate the model. Second, some relationships are, for whatever reason, intrinsically different across space. For example there might be different administrative, political, or other contextual issues that produce differing (people's) responses to the same stimuli across space. Third, a model form which the relationships are being measured can be a gross misspecification of reality. For example, one or more variables have been omitted from the model or represented by incorrect functional form.

A technique that deals with such spatial variation is GWR. GWR expands standard regression for use with spatial data. It allows the parameters to vary locally within the study area and might provide a more appropriate and accurate basis for descriptive and predictive purposes. It does not however allow extrapolation beyond the region in which the model was established (Foody, 2003 cited by Wang et al., 2005). GWR assumes spatial auto correlation and spatial non-stationarity. Spatial auto correlation is the situation at which a value of a variable at a location is related to the values of the same variable at the locations nearby. Spatial non-stationarity means that the relationships between independent and dependent variables are not constant over space (Tu & Xia, 2008).

GWR extends the traditional regression framework of equation:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \epsilon_i$$

with equation:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i v_i) x_{ik} + \epsilon_i$$

where y_i is the dependent variable, (u_i,v_i) are the coordinates, β_k (u_i, v_i) is a realisation of the continuous function β_k (u,v), and ε_i *is* the residual, all at point *i*.

The calibration of GWR equation assumes that observed data near to point *i* have more influence in the estimation of the β_k (u_i, v_i) than data located farther from *i*. A basis for understanding how GWR operates is provided by weighted least squares. In GWR, an observation is weighted in accordance with its proximity to point *i* so that the weighting of an observation is no longer constant in the calibration but varies with *i* (Fotheringham et al., 1998).

Any point of the regression depends not only on the observations received, but also on the choice of kernel and its bandwidth. Only independent variables, which are within the specified window around dependent variable, are taken into account when describing the relationship between the variables at a certain location. In general two kernel types can be used. A fixed spatial kernel assumes that the bandwidth at each regression point is constant throughout the study area. Alternatively, an adaptive spatial kernel adopts a variable bandwidth. In the latter kernel type, higher weights are assigned where data are more scattered while lower weights are assigned where data are denser or more abundant (Fotheringham, 2002 cited by Pineda et al., 2010). An example of this type of kernel is the bi-square function (Pineda et al., 2010):

$$W_{ij} = (1 - \frac{d_{ij}}{b})^z \quad si \;\; d_{ij} \leq b$$
$$W_{ij} = 0 \quad si \;\; d_{ij} > 0$$

where d_{ij} is the Euclidean distance between point *i* of the regression and the observed point *j*, and b is the bandwidth. If *i* and *j* coincide, the weighting of data at that point will be unity (1) and the weighting of other data will decrease according to weighting function (for example a Gaussian curve) as the distance between *i* and *j* increases.

Regardless of the weighting function employed, the essential idea of GWR is that for each point *i* there is a 'bump of influence' around it. Such a 'bump' is determined by the weighting function so that sampled observations nearby have more influence in the estimation of the regression parameters of *i* than do sampled observations farther away (Fotheringham et al.,1998). A graphic display of Gaussian kernel weighting function is shown on Figure 5. The regression is located at the point L_i; whilst W_{ij} is the associated weighting at the point located at L_j, d_{ij} is the distance between the point (of regression) L_i and the point L_j.

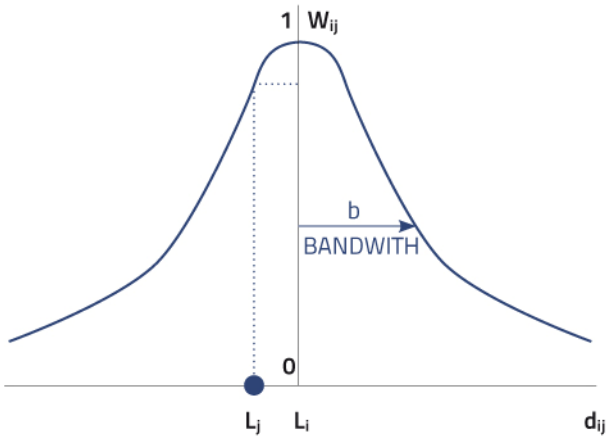


Figure 5. Gaussian kernel. Source: Fotheringham, 2002 cited by Pineda et al., 2010, p. 582

Fotheringham et al. (1998) acknowledges that a difficulty with GWR is that the estimated parameters are, in part, functions of the weighting function or kernel selected in the method. That is because as bandwidth becomes larger, the closer the model solution is to that of OLS and when bandwidth is equal to the maximum distance between points in the system, the two models will be equal. When bandwidth becomes smaller, the parameter estimates will increasingly depend on observations in close proximity to i and will consequently have increased variance. The problem therefore is how to select an appropriate bandwidth or decay function in GWR.

Collazos et al. (2006) (cited by Pineda et al. 2010) states, that it is always necessary to determine the optimum bandwidth when performing a GWR. It is possible to apply a bandwidth directly to the model if it is known a priori. In case it is unknown, which is usually the case, a cross-validation estimate can be applied, or alternatively, the Akaike Information Criterion (AIC) can be employed to help find the optimal settings. AIC is a measure of relative goodness of fit of a model which takes into account both its accuracy and complexity. The lower the AIC, the closer the approximation of the model is to reality (i.e. the best model is the one with the smallest AIC). As a rule of thumb, a 'serious' difference between two models is generally regarded as one in which the difference in AIC values between the models is at least 3 (Fotheringham et al., 2002 cited by Wang et al., 2005). AICc is AIC with a correction for finite sample sizes.

Important outputs of GWR are residuals, R^2 values and R^2 Adj values. The residuals are the parts of the dependent variable that are not explained by the model. The R^2 value is the coefficient of determination and the R^2 Adj, or adjusted R^2, bears more information as its values depend on a complexity of model as well. Finally, a cartographic analysis of the GWR results is important, as one can observe the distribution of high and low correlations according to their location and not merely according to their mean value (Bole, 2010).

According to Bole (2010) the use of traditional statistical methods (often applying only descriptive statistics) is still very common in spatial sciences, though spatial statistical methods such as GWR have been developed lately. There are various reasons for that. First, there is not much software available to use this method, as it has only been developed recently; one of the first authors that have set the foundations of application of GWR is Fotheringham (Fotheringham et al., 2002 cited by Bole, 2010). Second, GWR is more an exploratory than a confirmatory statistical method and thus much more suitable for applications such as data mining. Finally it requires a certain level of statistical expertise. Currently GWR is most commonly used in fields that in their core are not spatial, for example in economy and health related research. However with increasing availability of GWR enabling software, for example GWR 3 and GWR module in commercially successful GIS

software ArcGIS (from version 9.3 onwards), the application of the GWR method should increase in the field of spatial sciences as well (Bole, 2010).

In fact, the number of GWR applications in environmental or geography related sciences is growing. The comparison of GWR modelling capabilities to the capabilities of other modelling techniques is common. Many studies compare the GWR method to commonly used regression methods. For example, Wang et al. (2005), Tu & Xia (2008) and Bole (2010) compared the GWR method with the ordinary least squares (OLS) method. The goal of the first study was to obtain a net primary production model for forest ecosystems in China. The second study dealt with the examination of spatially varying relationships between land use and water quality. The third study analyzed settlements characteristics of the Ljubljana urban region. The results of all of these three studies indicate that GWR has the potential to become a useful tool for solving spatial problems as it has outperformed OLS. Also, GWR has shown local variations that remained hidden using the OLS technique.

The GWR technique was also applied in modelling yield of perennial plants in agriculture. For example, Perry et al. (2009) researched spatial variation in tree characteristics and yield in a pear orchard. They examined the spatial structure of fruit yield, tree size, vigour, and soil properties for an established pear orchard using several techniques, including GWR. In this research, GWR spatial significance tests failed at some scales so they were eventually unable to use GWR, though they acknowledged that GWR supports the conclusions of non-stationarity of the phenomenon. Baluja et al. (2012) researched a relationship between vine vigour and yield. More specifically, they focused on the spatial variability in anthocyanin (a pigment occurring in tissues of plants) content in grapes and quantification of its relationship with the vigour and yield. The location of research was a commercial vineyard in Spain. They discovered that application of GWR revealed considerable spatial heterogeneity in the relationships between grape colour with yield and vigour. These two researches can be regarded as an application of GWR in precision agriculture.

Furthermore, an extensive review of studies using bio spatial response variables that explicitly incorporate spatial dependence was done by Miller et al. (2007). They analyzed 54 studies and concluded that autoregressive methods, geostatistical methods, GWR and parameter estimation methods were applied in researches when dealing with spatial dependence. According to their research, GWR has so far been used mainly as a data exploration technique rather than a predictive method in biogeographical applications.

2.6 Self-organizing maps (SOM)

SOMs are in their core a clustering method and are used to both cluster the data and to visualize the relationships among the data items. SOM clusters are represented in a manner that preserves the non-linear relations and the topology of the data items. SOMs can be performed in an unsupervised or a supervised fashion. Unsupervised SOMs cluster the data based on all input variables. Supervised SOMs are applied when a dependent variable is available (Wehrens & Buydens, 2007).

The steps in SOM data mining are presented in Figure 6. The first step is data collection. The collected data can be regarded as raw data, as it is in most cases not appropriately prepared for the analysis. The aim of the next two steps, data preprocessing and data normalization, is to appropriately prepare the data for SOM training. The next step is data mining technique

application, SOM training, which results in various numerical and graphical outputs. These outputs are applied to visualize the results, to cluster the input data and to locally model input the data.

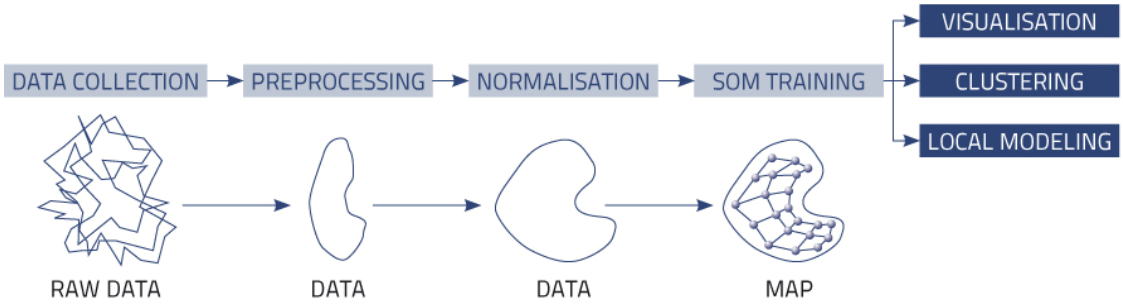


Figure 6. Applying the SOM in data mining. Source: Vesanto, 2002, p. 3

Teuvo Kohonen introduced the SOMs in 1981 (Kohonen, 1990). Extensive descriptions of the SOM algorithm are available in a number of literature sources (e.g. Kohonen, 1990; Weherns & Buydens, 2007; Malone et al., 2005). The brief description provided below is based on an outline of SOMs as presented by Henriques et al. (2012) and Vesanto (2002), both based on original SOM algorithm by Teuvo Kohonen.

The basic goal of a SOM is to map (project) the input data onto an n-dimensional grid of units, also referred to as neurons. Though this grid or map is usually 1 or 2 dimensional, higher dimensions are also possible although not often used because their visualization is problematic. The grid forms the output space while the original data space is the input space. The original SOM algorithm is based on unsupervised competitive learning, which means that the training (process of clustering) is entirely data-driven and that the neurons of the output map compete with each other.

Table 1 presents an example of input data. In this example, ecological data is presented in a matrix containing n species (Sp_1... Sp_n) that are observed in p sample units (SU_1... SU_p). Figure 7 shows the connection between the original data form the input layer (data from Table 1) with neurons in the output space (or output layer). Each sphere in the upper part of the figure symbolises an explanatory variable from the data matrix (input layer). In the lower part of the figure, each sphere represents a neuron of the output layer (SOM). The figure shows that a variable (in this example SU_j) is in a way connected to all the neurons of the output map, as data records of that particular variable contribute to properties of every neuron on the output map. In fact, each neuron's properties are derived from all variables of input data during the process of training.

Table 1. SOM input data example. Source: Giraudel & Lek, 2001, p. 330

		Sample units			
		SU_1	SU_2	...	SU_p
Species	Sp_1	X_{11}	X_{11}	...	X_{1p}
	Sp_2	X_{12}	X_{12}	...	X_{2p}
	...	...	...	...	...
	Sp_n	X_{n1}	X_{n2}	...	X_{np}

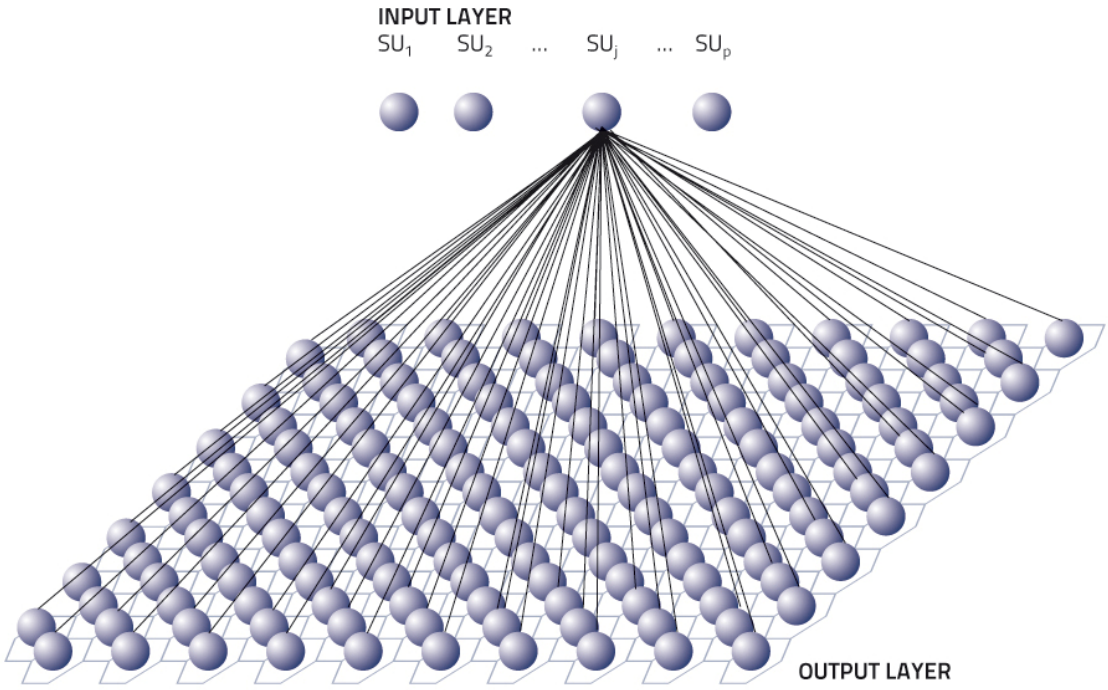


Figure 7. A two-dimensional SOM, based on Table 1. Source: Giraudel & Lek, 2001, p. 330

The process of training is as follows. First, the map (grid) properties (such as size, shape, lattice properties, etc) as well as the algorithm properties (such as neighbourhood distance, number of iterations, etc) are defined by the analyst. Before the training phase begins, initial values are given to the weight vectors. Weight vectors are actually the vectors which describe values of data records, where each vector weight is a normalized value of a particular dependent variable of that data record. The initialization follows one of the following three procedures: random, sample or linear. The training tries to preserve the topological relations, meaning that similar patterns from input space will be mapped to units that are close in output space. Each unit, being an input layer unit, has as many weights as the input patterns and can thus be regarded as a vector in the same space of patterns. When training a SOM with a given input pattern, the distance between that pattern and every unit in the network is calculated. The most common distance metrics for that purpose is Euclidean distance. In each training step, one sample vector from the input dataset is chosen randomly and a similarity measure is calculated between it and all the weight vectors of the map. Then the algorithm selects the unit that is the closest as the winning unit (also referred to as best matching unit – BMU), and maps that pattern (vector) onto that unit in output layer. After finding the BMU, the weight vectors of the SOM are updated. The weight vectors of the BMU and its topological neighbours are moved closer to the input vector in the input space. This adaptation procedure stretches the BMU and its topological neighbours towards the sample vector. The process iterates with a preset number of iterations.

The training is usually performed in two phases. In the first phase, a relatively large learning rate or alpha (α) and neighbourhood radius (r) (size of the neighbourhood around the winner unit in which units will be updated) are used. The alpha values and the neighbourhood radius decrease with increasing number of iterations. In the second phase (a.k.a. fine tuning), the starting alpha and neighbourhood radius values are smaller than at the beginning of the first phase and again decrease (towards zero). Alpha values are in the internal [0, 1] and must evolve towards 0 to guarantee convergence and stability in the training process. The evolution of alpha values usually follows a linear decreasing function.

The basic SOM learning algorithm may be described as follows (Henriques, 2012, p 220):

Let
X be the set of n training patterns $\mathbf{x_1}, \mathbf{x_2} \ldots \mathbf{x_n}$
W be a p x q grid of units $\mathbf{w_{ij}}$ where i and j are their coordinates on that grid
α be the learning rate, assuming values in [0,1], initialized to a given initial learning rate
r be the radius of the neighbourhood function $h(\mathbf{w_{ij}}, \mathbf{w_{mn}}, r)$, initialized to a given initial radius
1 Repeat
2 For m = 1 to n
3 For all $\mathbf{w_{ij}} \in \mathbf{W}$,
4 Calculate $d_{ij} = ||\mathbf{x_m} - \mathbf{w_{ij}}||$
5 Select the unit that minimizes d_{ij} as the winner $\boldsymbol{w_{winner}}$
6 Update each unit $\mathbf{w_{ij}} \in \mathbf{W}$: $\mathbf{w_{ij}} = \mathbf{w_{ij}} + \alpha\, h(\mathbf{w_{winner}}, \mathbf{w_{ij}}, r)||\mathbf{x_m} - \mathbf{w_{ij}}||$
7 Decrease the value of α and r
8 Until α reaches 0

As mentioned above, the analyst must select SOM characteristics before starting the training. As there is no obvious way to choose the properties of SOM for a certain dataset beforehand, the analyst may need to iterate a bit between training SOMs with different parameterization and validating them with different quality measures. There are several practical recommendations in literature that can help an analyst to prepare data and choose appropriate settings for SOM algorithm.

First of all, the quality of the input data is important because SOM learns directly from the data it receives. Following, results can be considerably affected by preprocessing the data (un)appropriately. Some basic preprocessing methods are component scaling, histogram equalization and filtering (Vesanto, 1997). Scaling of variables is of special importance if the Euclidean distance is used. If the data is not scaled, variables with higher values would dominate the map organization because of their greater impact on the calculated distance.

Properties of SOM map are important as well. The map size is important to detect the deviation of the data and if map size is too large it is possible to "overfit" the models (Céréghino et al., 2005). Vesanto et al. (2000) state that they usually use maps with 100 – 600 neurons for their explorations. Further, they have set the default number of neurons in a toolbox they created for Matlab software at 5 * sqrt (n) where n is the number of training samples. Finally, Vesanto et al. (2000) recommend the use of the hexagonal lattice, because then all 6 neighbours of a neuron are at the same distance (which is not the case in 8 neighbours in rectangular lattice as can be seen on Figure 8). Kohonen (1990) emphasised, that learning (training) is a stochastic process. Thus, the number of steps (iterations) must be relatively large to achieve accurate mapping. In this regard, he advises that the number of steps must be at least 500 times the number of network units (number of neurons in the output map). He also states that the number of variables has no effect on the number of needed iteration steps. Guidelines for initial phase choice of α and r values and for choice in fine tuning phase of training are given by Kohonen (1990) as well. Alpha (α) should start with a value that is close to unity for approximately first 1000 steps and then decrease to values of 0.01 or less over a long period. Finally, the initial radius (r) is often set to be more than a half of the diameter of the network and to decrease gradually to one unit.

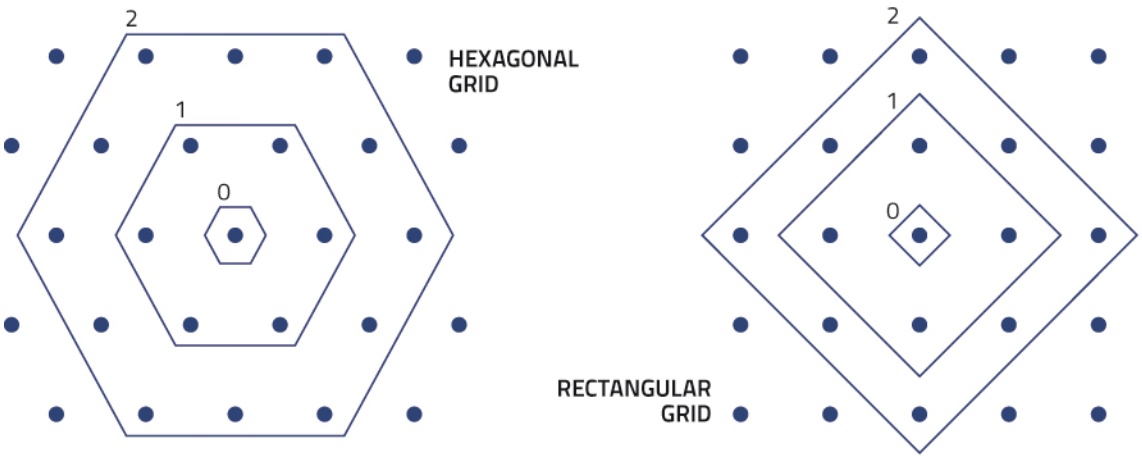


Figure 8. Hexagonal (left) and rectangular lattice (right). Source: Vesanto, 2002, p.12

The interpretation of SOM results is described in various literature sources (e.g. Vesanto (1997); Malone et al., (2005); ...). In Figure 9, one possible representation of an output layer of SOM is shown. The vectors in the circles represent the patterns of input data. One can observe that similar vectors are mapped close to each other.

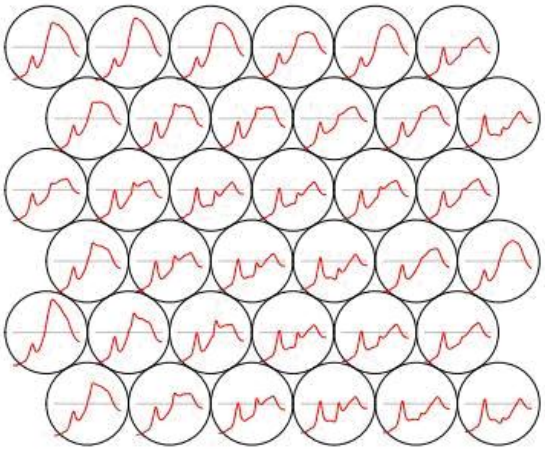


Figure 9. A possibility of visualization of prototype vectors in SOM. Source: Wehrens & Buydens, 2007, p. 10

SOMs need to be analyzed in details in order to derive information from the data. Two of the most information bearing SOM outputs are described and depicted below. The clustering structure of SOMs can be visualized by displaying distances between reference vectors; the most common method is the use of unified distance matrix (U-matrix) technique. The technique calculates the weight sum of all Euclidean distances between the weight vectors for all output neurons. The resulting values can be used to interpret the clusters created by the SOM (Malone et al., 2005). U-matrix can be shown in various ways (Figure 10), but it should be depicted in a way that the relative distances between adjacent map units on the whole map can be seen. Cluster borders can be identified as "mountains" of high distances separating 'valleys" of low distances (Vesanto, 2002).

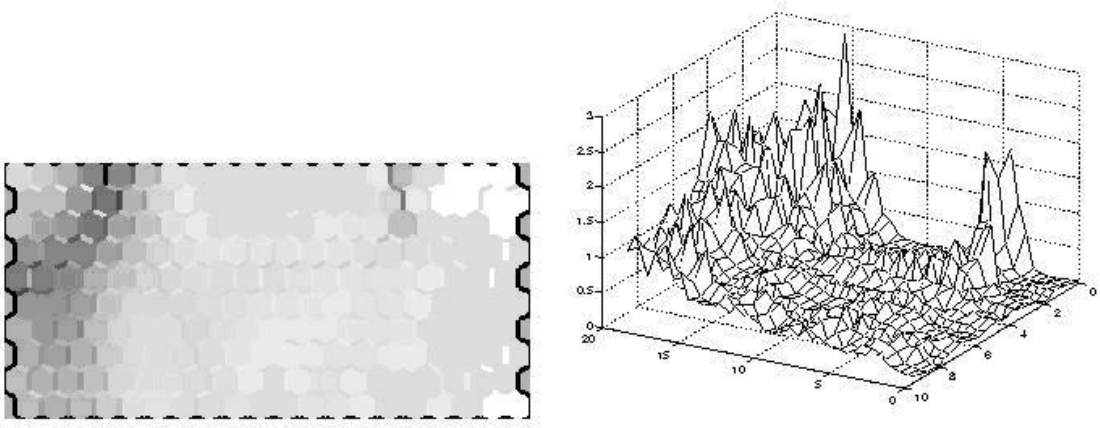


Figure 10. U-matrix in grey-level image (left) and as 3D plot (right). Source: Vesanto, 1997, figure 2.6

Another important analysis of the structure of SOMs is the visualization of their component planes. Component planes show the weights of the neurons of the trained SOM. 'In this technique the SOM can be thought of as a cake consisting of component layers. Each component plane is a horizontal layer of this cake while each reference vector is a vertical slice'(Vesanto, 1997, n. 15). Component planes are visualized by taking from each reference vector the value of the component and depicting this value as a colour according to colour scale. This kind of representation gives information about the distribution of the component values in the output layer of SOM. By comparing several component planes at the same time, one can easily notice simple correlations between components. An example is illustrated in Figure 11.

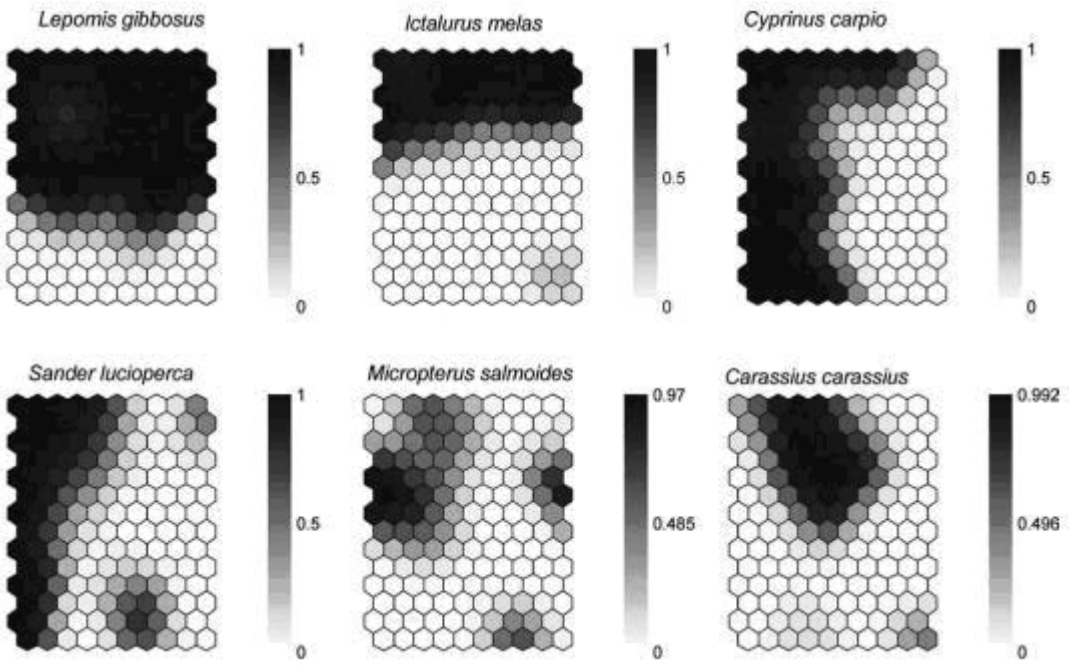


Figure 11. An example of component planes. Source: Céréghino et al., 2005, p. 463

The inventor of SOMs, Teuvo Kohonen, primarily used them in practical speech recognition. He analyzed the application of SOMs in other fields of research in the first decade after their introduction and exposed their use in robotics, process control and telecommunications (Kohonen, 1990). Though Kohonen's main interest was the application of SOM in speech recognition, he was involved in presenting possibilities of SOM usage in other fields, such as zoology (Ritter &

Kohonen, 1989) or economics & demography (Kaski & Kohonen, 1996). Nowadays SOMs are used in a number of scientific and commercial applications in fields such as medicine, biology, chemistry, image analysis, engineering, computer sciences and many more (Wehrens & Buydens, 2007). A bibliography of 7718 papers dealing with SOMs is currently available at the website of the university where Teuvo Kohonen initiated the SOM research (HUT, 2007).

Two interesting SOM applications relating to the subject of this thesis were found in the literature. A SOM data mining method was used by Shanmuganathan et al. (2010) to discover relations between daily extreme weather conditions and quantity of vineyard yield as well as quality of wine. They primarily applied χ^2 method, which they found in literature described within research dealing with similar problems they had had. The reason they searched for alternatives, which lead them to comparison of methods, was the lack of sufficient data in their particular case. Though only indirectly linked to our case, SOM data mining method was presented by Wehrens & Buydens (2007) as a possible method for clustering wine samples according to their chemical characteristics.

SOMs are usually applied for quantitative data exploration. However, qualitative data was also used. For example, Vesanto (1997) conducted a research dealing with paper mills using both quantitative and qualitative data. In his approach, he used a one-dimensional SOM in order to prepare qualitative data to include in his research, where he eventually used two dimensional SOMs. He however was able to classify each of the qualitative variables of his dataset to only a few classes. Ritter & Kohonen (1989) on the other hand used only qualitative variables in a SOM dealing with animal taxonomy. They used qualitative descriptions of animal properties (has feather, likes to run, etc) to create a table where they indicated presence or absence of particular qualitative property by a binary value 0 and 1. Vesanto et al. (2000) as well suggests such method if one needs to utilize symbolic variables in training the SOM.

As the number and diversity of fields in which SOMs were applied grew, a number of algorithm variations as well as computer software applications were made available to suit different purposes. For example, staff from the Helsinki University of Technology developed a package named 'SOM Toolbox", which can be used in Matlab software. C-code can be obtained from the same source as well (Vesanto et al., 2000). Further, three packages that support SOM ('som", 'class" and 'kohonen") are available for the freely available R statistical software (Wehrens & Buydens, 2007). Finally, an interesting suite from the geographical perspective is GeoSOM. It enables training SOMs with an ordinary algorithm as well as with a modified algorithm, which takes geographical information into account when training SOM (its weight in training is defined by user) (Henriques et al., 2012).

3. STUDY AREA AND DATA

First two subchapters of this chapter focus on the study area. Basic facts about Slovenia, particularly from the perspective of viticulture, are presented first and are followed by the description of the Goriška Brda wine district, which is the study area of this research. The reasoning for the choice of this particular research area is given as well.

The remaining subchapters focus on data. First, the data that derives from farmers declarations are described. That is GERK (graphical unit of farm holding) data, vineyard register data, and harvest and production declaration data. 'Vineyard register data" and 'harvest and production declaration data" are formally parts of the 'Register of grape and wine growers", but are here described separately for clarity of presentation. These datasets are administrated by the Slovenian Ministry of Agriculture and Environment (MAE), while the data are mostly gathered for the purpose of subsidies, taxes and for monitoring. Second, the meteorological data used in the research are described. Finally, the data that was acquired for the purpose of the research, but was eventually not used, are briefly mentioned.

3.1 Slovenia

Slovenia is a county in central Europe, bordering Italy to the west, Austria to the north, Hungary to the east and Croatia to the south (Figure 12). It covers 20 273 square kilometres and has a population of just over 2 million. The country includes part of the Alps in the northern and north-western part, Panonic Basin in the eastern part and the Dinaric Mountains in the southern part of the country. It has access to the Adriatic Sea in the south-west. Slovenia has a very dynamic relief with few flat areas, mostly in the eastern part. Consequently most of the area of the country is classified as less favoured for agriculture by EU standards.

Nevertheless, Slovenia was known for its good conditions for growing vine in past, as the presence of vine was first documented 2600 years ago. Climatic conditions are important from that aspect. The climate of the country is influenced by the Alps, the Mediterranean Sea and the Panonic Basin. With annual average temperatures between 9 °C and 13.7 °C in viticultural areas, production of diverse varieties is possible (Kuljaj, 2005). There are three wine growing regions in Slovenia, Podravje, Posavje and Primorska (Figure 12), which are delimited according to ecological conditions, wine characteristics and some other factors like tradition (Uradni list, 2003).

Vineyards in Slovenia are mostly planted on southern, south-eastern and south-western slopes of hilly areas, as climatic conditions in most of the flat land are not suitable for growing vine. Also, the growing conditions in certain areas do differ significantly on micro scales as well and are a factor in yield quantity and quality (Vršič & Lešnik, 2001). Because natural conditions in which vines are grown differ, the selection of grape varieties by wine regions and by micro locations differs as well (Rusjan & Korošec-Koruza, 2003). In fact, farmers are only allowed to grow recommended grape varieties in certain wine regions or sub-regions (Uradni list, 2003).

In Slovenia, viticulture is important because the majority of the wine produced in the country is high quality wine. Additionally, viticulture in Slovenia is unique because of the steep slopes and small average size of the vineyards. This results in high numbers of "small" vine-growers, many of which produce wine mostly for self consumption (Jakša, 2011).

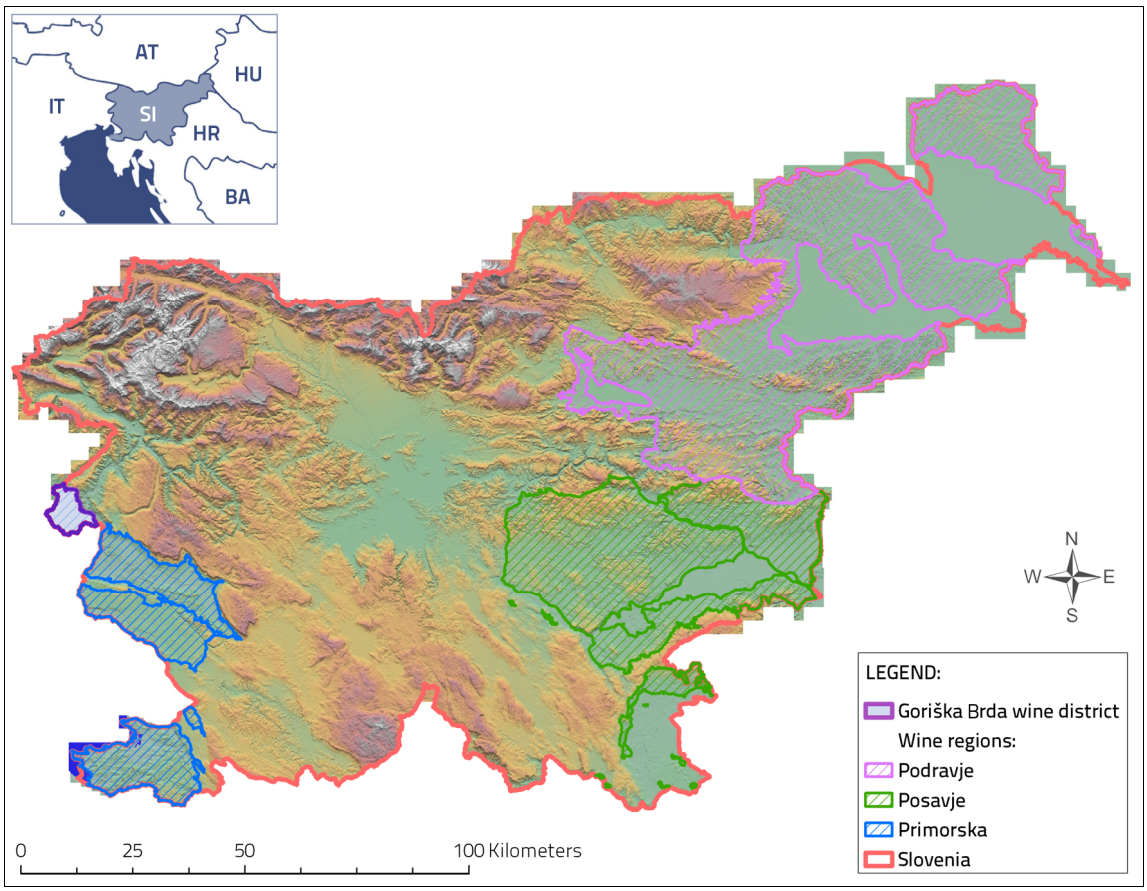


Figure 12. Wine regions of Slovenia and country location in central Europe. Source: MAE (2012a), MAE (2012b)

3.2 Goriška Brda

The Goriška Brda wine district is part of the Primorska wine region, which is located in the western part of Slovenia (Figure 12). This wine region is climatically influenced by the Mediterranean Sea. Consequently it has the warmest and sunniest climate of the three wine regions of Slovenia. It has more than one third of Slovenia's vineyards and produces more than two-fifths of its wine (Kuljaj, 2005). The region is divided into four wine districts, out of which Goriška Brda wine district is the northernmost and the most relief intensive one.

Goriška Brda stretches over a hilly terrain (brda means hills in Slovenian), covering the area of 66 square kilometres. Though altitudes within the area vary from 50 to 600 meters above sea level, all of the declared vineyards are located at altitudes below 430 meters. The altitudes, which increase towards the north, are the lowest on the outskirts of Frullian Lowlands in the south (Figure 13). The area is orographically opened to south/south-west towards Frullian Lowlands and further towards the Adriatic Sea, which is located approximately 20 kilometres to the south.

The vicinity of the Mediterranean Sea results in mild winters and warm summers, which is ideal for vine. The average yearly temperature in Bilje (Figure 16) is 11.8 °C. The warmest month is July with average temperature of 21.4 °C, though the average August temperature is over 20 °C as well. The average January temperature is 2.7 °C (SEA, 2012). The most common winds are south-western winds, supplying warm and humid air. The level of precipitation increases towards the north as it follows the increase in altitude. For example in the lowest altitudes, the average yearly precipitation is 1456 mm (Bilje, 10 km to SE, 55 meters altitude) (SEA, 2012), while at somewhat

higher altitudes (Vedrijan, 241 meters altitude) the average yearly precipitation is 1704 mm (Zupančič, 1995). Precipitation is also temporally unevenly distributed. The precipitation (in Vedrijan) is the highest in June (176 mm) and in November (172 mm) and the lowest in February (96 mm). After a minimum in February, the level of precipitation steadily rises until June and usually provides enough reserve water supply to avoid drought in July and August, when the temperatures are highest. Though vine yield can be affected by lack of water in drought years, usually only areas on high slopes of Goriška Brda might be affected by some drought due to thin soil. The most dangerous weather related risks for vine in Goriška Brda area is the Bora wind. This wind is caused by high air pressure in the hinterlands, and springtime frosts, occurring in case of lengthy occurrences of cold winds during anticyclone (Belec et al., 1998).

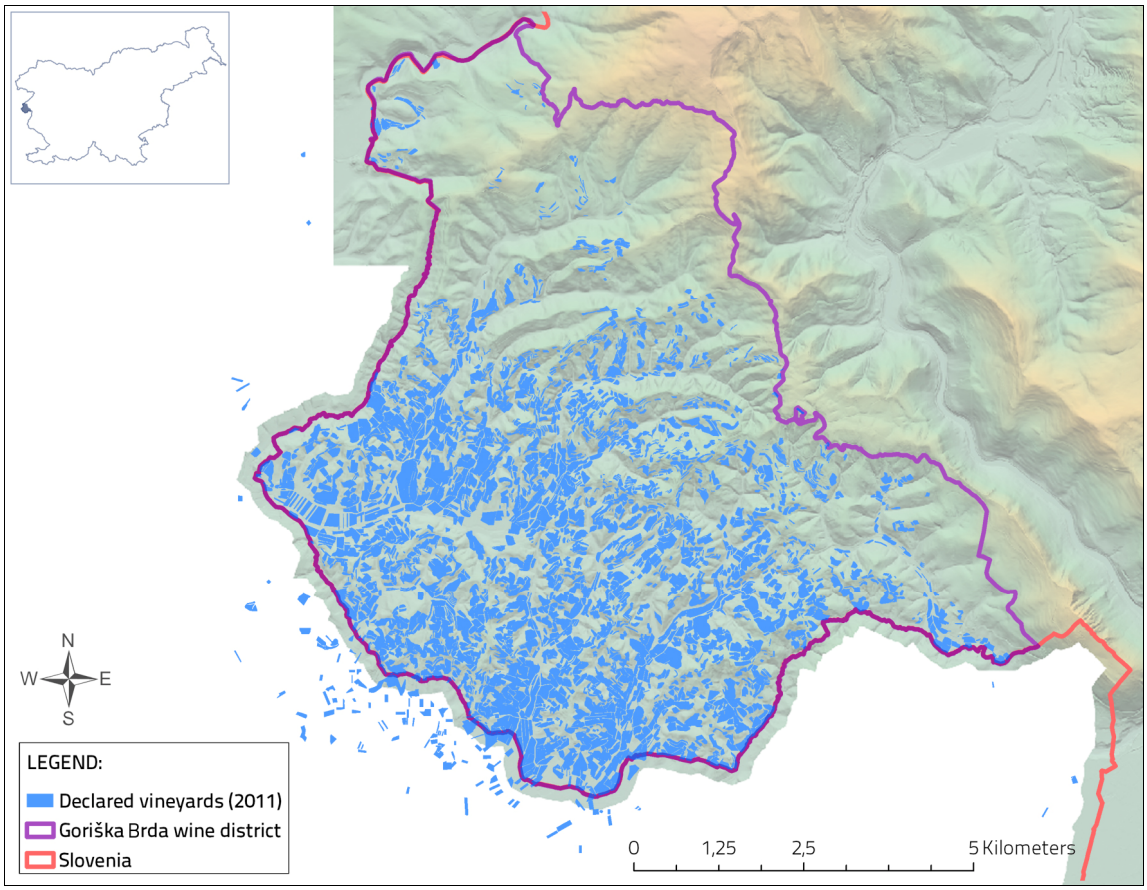


Figure 13. Declared vineyards in Goriška Brda study area in 2011. Source: MAE (2012a), MAE (2012b)

The soils in the Goriška Brda region are composed of ocean sediment mass, as layers of flysch, sandstone and limestone remained after the sea disappeared. Flysch weathers fast and changes into fertile soil suitable for growing wine (Belec et al., 1998). However, the soil is poor in organic and mineral substances, phosphorous, potassium and other nutrients (Drnovšček, 1994). This can negatively affect the vines nutritional needs (Kuljaj, 2005). Nevertheless, vine is the most important crop in Goriška Brda.

Recommended wine varieties in Goriška Brda are Rebula, Sauvignonasse (Tocai friuano), Pinot blanc, Sauvignon blanc, Malvazija, Pinot gris, Chardonnay, Merlot and Cabernet Sauvignon. Allowed varieties in the area are Muscat Blanc, Pikolit, Prosecco, Verduc, Refosco, Cabernet Franc, Pinot Noir, Barbera, Syrah, Gamay, Glera, Klarnica, Pergolin, Pokalca and Poljšakica (Uradni list, 2003). Some of the allowed varieties are local, grown in small quantities and important primarily as local speciality. The district is well known for its Rebula and Sauvignonasse (Tocai friuano) wines, Pinot Blanc and Chardonnay also reach excellent quality, while Merlot and Cabernet Sauvignon are

principle red wines (Kuljaj, 2005). The number and ratio of planted vines of most common varieties in Goriška Brda, as declared by farmers in 2011, is presented in the Table 2.

Table 2. Most common varieties in Slovenian part of Goriška Brda as declared by farmers in 2011. Source: MAE (2012a)

	Number of vines	Ratio (to overall)
Rebula	1 340 690	20.55 %
Merlot	1 190 034	18.24 %
Chardonnay	1 139 422	17.47 %
Pinot gris	732 857	11.23 %
Sauvignonasse (Tocai friuano)	557 297	8.54 %
Other (23 varieties)	1 563 377	23.96 %
Overall	6 523 947	100 %

Figure 14. Vineyards in Goriška Brda near Vipolže.

Out of 1950 ha of vineyards in Goriška Brda (2009 and 2010 ortophoto interpretation), 1850 ha were declared by farmers in 2011. The ratio of declared area of vineyards is thus 95% which is the highest ratio of declared vineyards among all wine districts in Slovenia. The distribution of declared vineyards in Goriška Brda is shown in Figure 13.

Out of a number of wine growing regions in Slovenia, Goriška Brda was chosen to illustrate this research because of a number of reasons, mostly as they are connected to the goal of finding

relationships between natural and socio-economical characteristics with yield. The reasons in favour of choosing Goriška Brda as the study area of this thesis are:

- It is a hilly area with vineyards on various expositions, slopes and altitudes,
- It has a high percentage of declared vineyards, which increases the possibility of relating yield data to exact location (see sections 4.1 and 5.1 for details),
- It has a high percentage of declared yield, which increases the possibility of relating yield data to exact location (see sections 4.1 and 5.1 for details),
- It is an area known for wine making and high quality wines and therefore there is a good possibility that farmers declarations (estimations) of yield are correct above the average,
- Vineyards are cultivated by farmers with one or a few vineyards and by farmers with many vineyards, thus socio-economical differences might be present,
- It has a precipitation measurement station in the area of research and meteorological station close to the area of research (both by Slovenian Environmental Agency - SEA),
- There are several meteorological stations by Phytosanitary Administration of the Republic of Slovenia (PARS) in the area,
- The area is not irrigated (though vineyards in Slovenia are very rarely irrigated), so precipitation data can be used without restraints.

There are however some drawbacks pertaining to the choice of the area:

- Because of a relatively high number of vineyards per farm (3,75 on average), there is a high possibility that many records for most popular varieties will be useless as yield can not be traced in exact location in certain circumstances (see sections 4.1 and 5.1 for details).
- Because of scarcity of meteorological stations and hilly relief, interpolation of meteorological data can present a problem.

In general, the Goriška Brda wine district has a good ratio between positive facts and drawbacks when comparing it to other vine growing areas in Slovenia from the perspective of yield estimation research. Therefore, it is a suitable area to test a new method, which could be applicable elsewhere in Slovenia in case it proves feasible in this area.

3.3 Vineyard polygons (deriving form GERK data)

GERK stands for Graphical unit of farm holding (in Slovenian). It is the basic unit of the Slovenian Land parcel information system (LPIS). According to legislation it includes a continuous area of same agricultural use, cultivated by one farm holding (Uradni list, 2006). Usually, one farm holding has many GERKs, representing their declared agricultural land. Farm holding can be registered by an individual or by a private company, though the vast majority of farm holdings are registered by individuals. The GERK registry was introduced in 2005 for EU subsidy related issues.

GERK data are acquired from farmer's declarations. The procedure of declaration is as follows. A farmer declares the land he cultivates to the administrative worker at the administrative unit where his/her land is located. The administrative worker, in the farmers presence and based on the farmers instructions, digitizes GERK polygons using a customized GIS application. Ortophoto images of various years, as well as some other layers (e.g. cadastre, land use ...), are available for them to help to distinguish boundaries of agricultural use. A farmer is responsible to update his GERK declarations in case if (s)he makes changes on the field (e.g. different boundaries, different land use ...). Following, the temporal and spatial accuracy of GERKs depends on the farmers'

interest in accuracy, on how up-to-date the ortophotos (0 – 3 years) are and on the skills of the administrative workers. However, as vineyards are easily distinguishable on ortophotos, GERKs' spatial accuracy for vineyard GERKs is generally good. Also, because vineyards are linked to obligatory declarations in some other registers (described below), their temporal accuracy is usually good as well. Some other attributes of GERK (given below) are calculated by software, and therefore correct. Finally, there are no topological errors, multi polygons, overlaps or other common errors in polygon data, as the system does not allow them. Overall, the quality of this dataset is good, especially from the technical aspect.

Farmers in Slovenia are obliged to declare their vineyard(s) to GERK registry if their total area exceeds 500 m^2 or if they sell grapes for wine production. Though the ratio of declared vineyards is increasing, there are still some undeclared. However, some vineyards will never have to be declared if the threshold for obligatory declaration will not change considerably. In 2011, there were approximately 3200 vineyards polygons, declared by approximately 850 distinct farm holdings in the study area. The average size of a vineyard polygon was approximately 5900 m^2. GERKs in the area of research are shown in Figure 15.

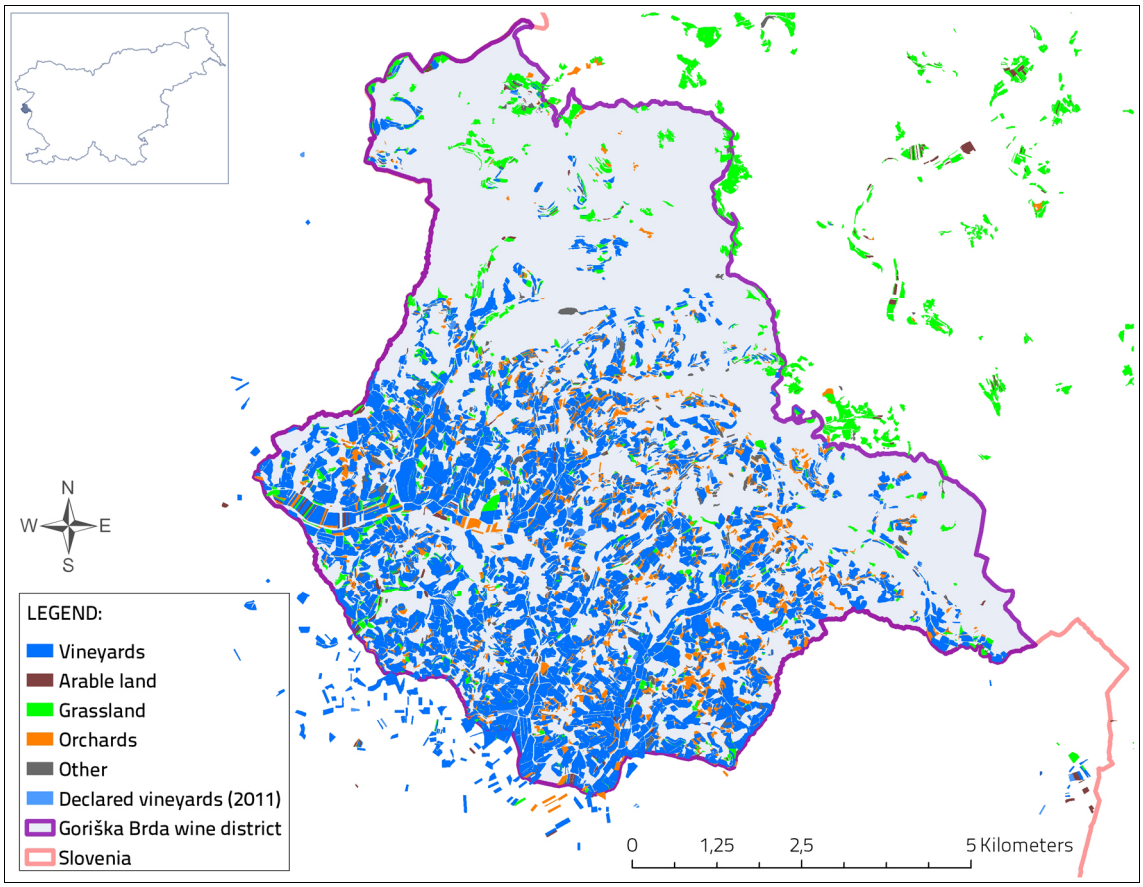


Figure 15. GERKs in Goriška Brda wine district in 2011. Source: MAE (2012a)

Technically, GERK is a georeferenced polygon. Besides attributes provided by a farmer, some of GERKs attributes derive directly from the polygons geometry, while some others are derived from intersection with digital terrain model (DTM) of 5x5 meters. As MAE does not possess DTM of Italy, only the GERKs located in Slovenia have all the attributes. Following, the polygons in Figure 15 that are outside the border of Slovenia can not be taken into account for this research. The attributes that are used in the research are presented below. In case the attribute is an explanatory variable in the research, its abbreviation as used further in this document is in brackets:

- Land use – only vineyards are used,
- GERK PID - unique GERK ID,
- Farm ID - unique farm ID,
- Area (PARC_AREA),
- Altitude - average altitude of polygon according to DTM (Z_AVG),
- Exposition - average exposition of polygon according to DTM (EXP_AVG),
- Slope - average exposition of polygon according to DTM (SLOPE_AVG).

The GERK data used in this thesis corresponds to that of the 30th of June each year. This date was chosen because of two facts. First, an earlier date could provide an outdated GERK as the majority of the updates occur before or during the subvention campaign, which usually spans from February to mid-June. Even if a farmer makes changes on the field in winter or autumn, they would usually not declare the changes until the next subvention campaign. Second, a couple months later the data could include changes in GERK that would occur after the harvest, which would not be acceptable, because the yield as declared in a certain year (dataset described below) is grown on plants that have been growing in vineyard at the time of harvest.

3.4 Vineyard register data

The vineyard register data contains vineyards characteristics as declared by the farmers. The declaration of vineyard characteristics in digital form was introduced in 1999. From 1999 to up to 2006, vineyard characteristic were declared based on cadastral parcels. In 2007, the system underwent a second major change, as existing attribute data was linked with GERK (abandoning cadastral parcels). Since 2007, vineyard characteristics are declared based on GERK (Jakša, 2011).

Farmers are obliged to declare vineyard characteristics as well as to update the data in case of changes. A farmer declares characteristics (or their changes) during a meeting at the administrative unit or by sending a written form. Even though attributes in this register are supposed to be correct (as applied by the farmer), there is a possibility of slight error in one attribute in particular: the number of vines. This attribute is in some cases not sufficiently up-to-date. Farmers, in general, promptly declare new wines when they restructure a vineyard or its part. It is however possible that some farmers do not promptly declare the decrease of number of vines, which sometimes occurs due to disease or meteorological extremes. The data might also be wrong in case of false declarations or unintended mistakes, but the data's accuracy in certain fields is automatically cross checked with data form GERK dataset during their input. For example, number of plants is compared with GERK size, certain time frame is allowed for selection of year of planting, etc. Input is rejected in case of extreme deviation (Jakša, 2011).

Vineyard characteristics can be easily linked to GERK register via the GERK PID (unique GERK id). The attributes from vineyard register that are used in this research are listed below. In case the attribute is an explanatory variable in the research, its abbreviation as used further in this document is in brackets:

- GERK PID,
- Farm ID ,
- Grape variety,
- Number of vines (PLANT_SUM),
- Year of planting (Y_PLANTING),
- Row spacing (R_SPACING),
- Vine spacing (V_SAPCING).

For consistency, the date selected to collect/extract the vineyard register matches that of the GERK registry (i.e. 30th of June of each year).

3.5 Harvest and production declaration data

The harvest and production declaration data contains information about grape yield (yield) and wine production. Data are declared by the farmers or by the companies that buy their grapes. According to Slovenian legislation, farmers are obliged to declare their yield, if the sum of the whole area of their vineyards exceeds 1000 m^2 or if they sell grapes for wine production (Uradni list, 1999b). The farmers are required to declare yield (in Kg) and wine production (in L) separately for each grape variety and quality type they produce. An important fact is that the yield is declared at farm level (not at GERK level), which presents a problem from the perspective of this research and will be elaborated further in methodology chapter. Finally, though harvest and production has to be declared by 20th November every year, some farmers declare it even later, as some types of wines require very late harvest (Jakša, 2011).

The yield quantity is based on farmers estimations. Farmers are not required to actually weigh the harvested grape, they simply estimate it by considering the amount of wine produced that year (Jakša, 2011). Considering the data collection methodology, one can conclude that such estimations can be inaccurate (or even deliberately wrong).

An important fact is that some farmers do not declare their yield at all, because they have small area of vineyards and produce wine only for their own consumption. In these cases the legislation allows them not to declare their yield. However, according to other MAE datasets and field inspections, there are some farmers that do not declare yield, though they are obliged to. The ratio of such farmers is decreasing, mainly due to inspections on field (Jakša, 2011).

The following attributes of declared yield data will be used in the research:

- Farm ID,
- Grape variety,
- Declared yield in kg.

The data used in this thesis are the data available in the MAE databases on the 23rd of February of 2012 and includes all yield declarations from 2007 to 2011. That means that even the possible late declarations of yield harvested in 2011 are included.

3.6 Meteorological data

The Slovenian Environment Agency (SEA) measures meteorological characteristics using a network of meteorological stations. The meteorological station, located closest to the area of research, is located 10 kilometres south east of the southernmost part of the study area. This station is located in a plain which has a similar altitude (Bilje, 55 meters altitude) as the lowest, southern parts of the study area (50 meters altitude). The location of this station, as well as the locations of other stations, whose data was not used in this research, can be seen in Figure 16. The argumentation of why only this station's data was used for the research is given in section 5.1.

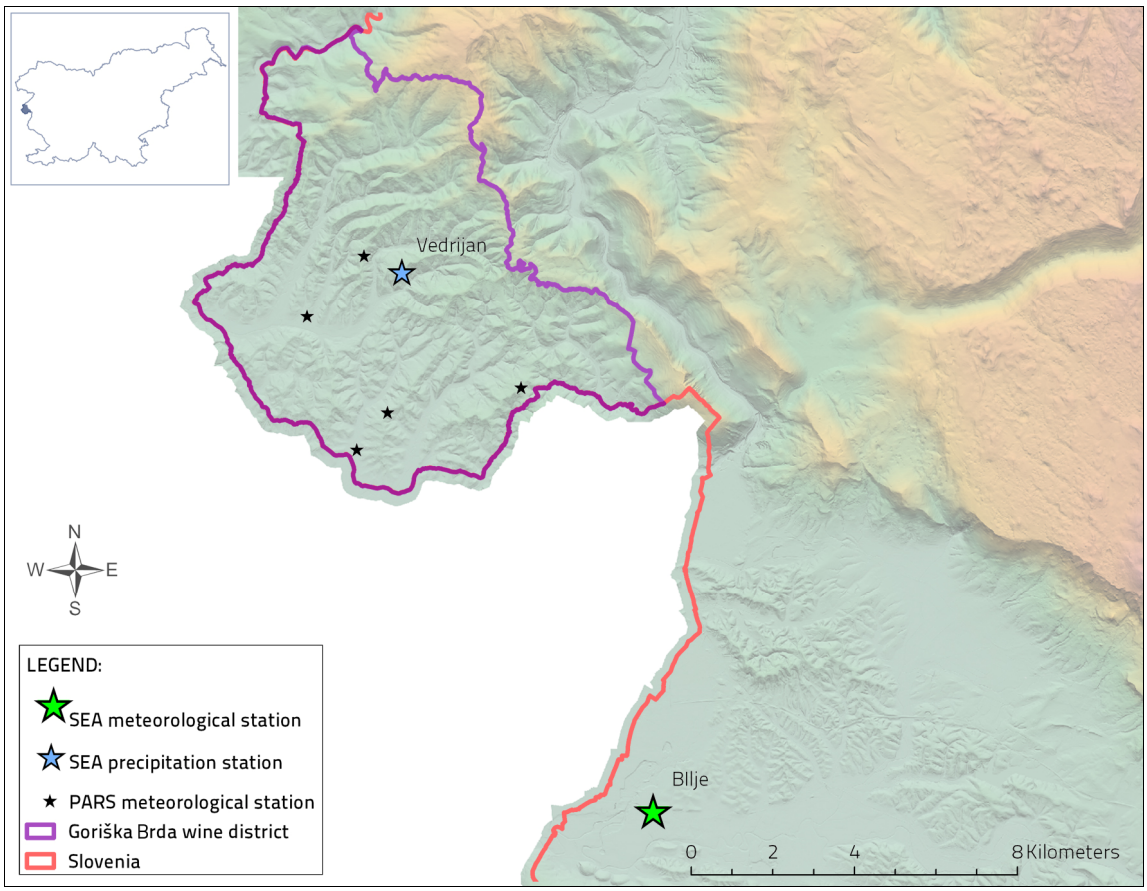


Figure 16. Location of meteorological stations in (or near) the area of research. Source: SEA (2012), PARS (2012), MAE (2012a), MAE (2012b)

The quality of the data is high because the meteorological phenomenons are measured by sensors. The problem, for this research, is the scarcity of measurement stations, as we can only use data from a single station, located outside the area of research. The scarcity of stations is particularly problematic because of the intensive relief and high difference in altitudes in this area. Thus, some available and potentially important measurements from the meteorological station (such as average wind speed ...) will not be used for this research. Furthermore, one must be aware that the representativeness of the measurements used for this research generally drops with increasing distance and an increase in altitude difference between the meteorological station and the vineyards.

The following data will be used from meteorological station (Bilje):

- average daily air temperature at 2 m (° C),
- 24-hour rainfall at 7 h (mm),
- number of sun hours.

The data mentioned above is gathered daily and is complete for the whole period from 2007 to 2011. Monthly means (temperature) or monthly sums (rainfall and precipitation) as well as yearly means and sums will be used for the research.

3.7 Other data

The following data was acquired as well because, according to literature, it could explain variation in vineyard yield:

- meteorological data from a network of stations by PARS (Phytosanitary Administration of the Republic of Slovenia),
- (detailed) soil map of Slovenia.

Furthermore, some of the variables in the vineyard register data were eventually not used, though they were acquired and initially intended to be used as explanatory variables. Vine training system type and rootstock type were not used at all. Grape variety type on the one side was used to identify Rebula variety records, but on the other side not used as an explanatory variable though initially intended as well.

This data was not used in the research due to reasons described in sections 5.1 and 5.4.

4. METHODOLOGY

The description of the methodology is divided into five parts. The first part describes data preprocessing, because most data preprocessing was common to all methods. The next three parts respectively present the research methods used in this research: OLS, GWR and SOM. The SOM part is somewhat more detailed, because SOM are in its core not a prediction method and because its testing, application and review is more complex. Finally, the fifth part describes the methods for comparing OLS and GWR accuracy results to meteorological data.

This research included many attempts during the data preparation phase and during data exploration phases, which did not yield desired results. Consequently, much of the acquired data were eventually not used to produce final results. Attempts to include that data in the research are briefly mentioned in discussion in sections 5.1 and 5.4. Only those research steps that led to the results presented in sections 5.2 - 5.5 are described here and illustrated in Figure 17 below.

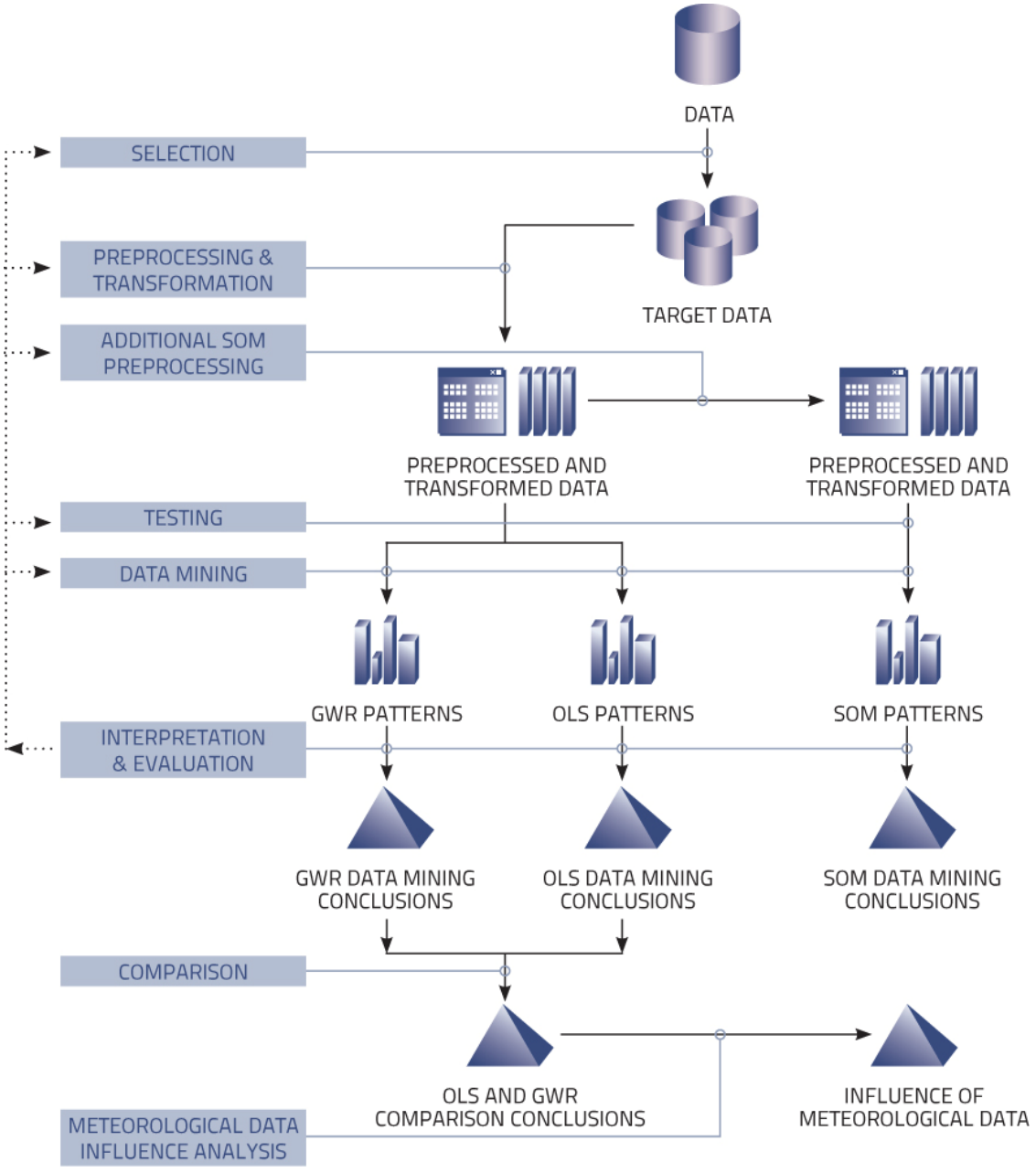


Figure 17. Research workflow.

Figure 17 presents the main steps of this research. The data acquisition was followed by data selections. Second, the data preprocessing that was common for all methods was performed (described in section 5.1). However, additional preprocessing had to be performed for data for the SOM method. Then the methods were applied (described in sections 5.2 – 5.4) using the same MAE data explanatory variables. These steps required a number of iterations to optimize the parameters and the results. The results of GWR and OLS were compared to detect the more accurate method of the two. The results of the SOM method were reviewed to asses a suitability of this alternative method for our research. Finally, the influence of SEA meteorological data on accuracy of GWR and OLS results was reviewed (described in section 5.5). All cartography was done using ArcGIS software.

4.1 Data preprocessing

A large part of data preprocessing was common for data used in all methods. Data preprocessing was required due to a need for proper preparation of data for application in data mining and due to the fact that new properties or variables had to be derived from raw data. The common data preprocessing can be divided into two main steps: data joining and data manipulation. Meteorological data used in OLS and GWR were prepared as well. Overview of data preprocessing is presented in the Table 3 below.

Table 3. Data preprocessing steps in the research.

	Input dataset	Actions taken within this step	Output data
Data joining	- vineyard centroids - vineyard characteristics - declared yield	- exclusion of records of which yield could not be traced to exact location of origin or joined with proper data records - joining of MAE datasets	- joined MAE data with location, properties and yield information for every data record
Data manipulation	- data joining output	- calculation of dependent 'yield per vine' variable - calculation of additional explanatory variables (socio-economic etc.) - log transformation of certain variables - detection and removing of outliers - multicollinearity checks	- joined MAE data with additional variables prepared for data mining
Meteorological data preparation	- SEA meteorological data	- calculation of yearly and monthly sums and means	- preprocessed SEA data

- Data joining

Joining of the datasets had to be performed in order to derive information from separate MAE datasets. It was performed in R statistical software system (R, 2012), using "plyr" and "foreign" packages. This step included three datasets, namely vineyard polygons, vineyard characteristics and declared yield datasets. The data records for which the yield could not be traced to exact location of origin were excluded. The data records for which yield could be traced to location of origin were kept and joined. The UML class diagram, showing attributes used in the research from raw MAE data and a relation between these three datasets is in Figure 18.

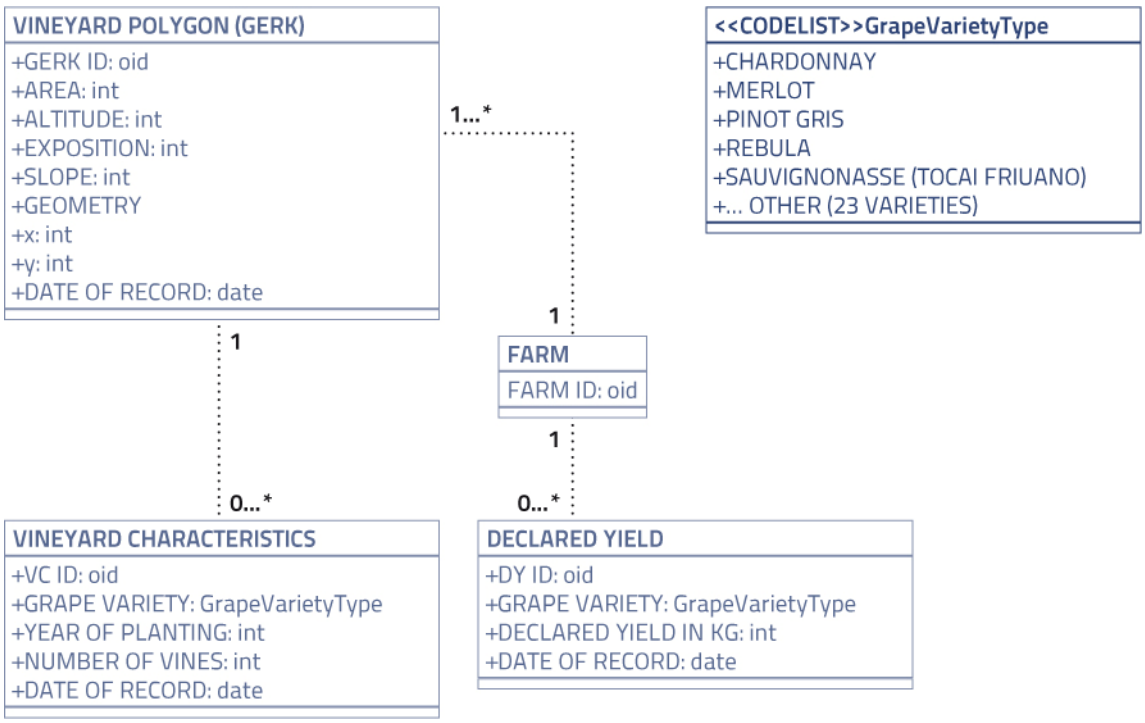


Figure 18. UML class diagram of MAE data used in the research. Source: MAE (2012a)

The diagram shows, that a farm in our case can have one or more vineyard polygons, each of which can have zero to many vineyard characteristics records. In case one vineyard polygon has more than one vineyard characteristics record of the same grape variety (for example different planting year), none of the records can be used for the research, because vineyard characteristics are explanatory variables (and we do not know how much yield is derived from which vines). One can also see in the diagram that a farm can have zero to many yield declarations, but the yield is declared separately only for each grape variety and not for each of its locations. A farm can therefore have a certain grape variety planted in many vineyards, but declares yield for that particular grape variety only in one record. Consequently, such yield can not be traced to a single vineyard polygon, and these data records can not be used for our research. The kinds of records mentioned in the examples above were excluded during the data joining phase of preprocessing.

- Data manipulation

Data preprocessing included data manipulation tasks, applied to data joined in the previous step. After new variables were calculated from existing ones, some of the variables were log transformed, outliers were detected and removed and finally multicollinear variables were detected and removed.

New variables were calculated in order to first, derive the dependent variable for the research, second, calculate a more exact area variable and third, calculate socio-economic variables. They are listed below (abbreviation as used further in this document is in brackets):

- yield per vine in kg - the dependant variable in the research, calculated from suitable declared yield and vineyard characteristics records (YI_PER_PLA),
- net area of grape variety with the same characteristics in vineyard - based on 'number of plants", 'vine distance" and 'row distance" (NET_AREA),
- area of all farm's vineyards - a socio-economic variable (F_AREA_SUM),
- sum of all farm's vine plants - a socio-economic variable (F_PLANT_SUM).

Log transformation of the following variables was required because their distributions were extremely skewed:

- sum of grape variety plants in vineyard (PLANT-SUM),
- area of vineyard (PARC-AREA),
- net area of grape variety with the same characteristics in vineyard (NET-AREA),
- area of all farms vineyards (F-AREA-SUM),
- sum of all farm's vine plants (F-PLANT-SUM).

Outliers were searched for in explanatory variables and in the dependent variable. The outliers in the explanatory variables (Z_AVG, R_DIST) were detected using visualizations of explanatory variables. Based on expert knowledge and a range of values depicted in visualizations, certain data records were removed from the dataset. However, the outliers in Y_PLANTING explanatory variable were detected based on literature (Vršič & Lešnik, 2001), according to which firstly, vines do not produce yield in the first year and secondly, vines are usually replaced after yield ratios decrease. Following, the data records of 1 year old vines and more than 50 years (due to different variety characteristics) old vines were removed.

When searching for outliers in the dependent 'yield per vine" variable, different approach was applied. The data records with less than 0.5 kg or with more than 6 kg of yield per vine were identified as outliers. The lower and upper limit for identification of outliers was set according to two facts. First, according to various sources (Martin, 2012; Abbazia, 2012; Stegovec, 2012) even viticulturists who deliberately decrease their yield per plant by cutting, still usually harvest more than 1 kg of grapes per vine. Second, according to Slovenian legislation, only grapes deriving from plans, that have up to 3 kg of grapes per plant, can be used to make wines labelled as quality wines (Uradni list, 2004). Because of the fact that the region is known for its high quality wines, setting of the limit twice as low / high as attempted / allowed, has been considered as reasonable.

Multicollinearity between the variables was identified by calculating and reviewing Pearson's correlation matrices for 2007 – 2011 data and for 2007 – 2011 Rebula data. VIF was calculated and its results were compared to Pearson's correlation matrices. Out of pairs of explanatory variables identified as collinear by VIF, the explanatory variable that had the largest correlation with the dependent variable (in majority of years of 2007 – 2011 period) was kept for further research.

- Meteorological data preparation

SEA meteorological data from Bilje meteorological station were selected among SEA and PARS meteorological stations data for research for a number of reasons (mentioned in section 5.1 as well). Station Bilje (60 m altitude), though it lies outside the area of research, is the only official national meteorological station in or close to the area of research and is therefore the most reliable. Also, it is the only station without missing data. Finally, it is the station that measures certain phenomenons that are not measured by any other station within the area. Eventually, the following meteorological characteristics were calculated:

- Precipitation sum,
- Sum of sun hours,
- Average temperature.

In these calculations, the data from including April to including September was used. The reason for this particular period comes from one of the approaches of calculation of viticulture climatic coefficients. Such coefficients can be used to help to determine a terroir; a wine growing region

with traditional viticulture and with characteristic climatic conditions (Rusjan & Korošec-Koruza, 2003). In our case, the selection of period of meteorological data was based on Huglins heliothermic index, which considers mean temperatures from April to September as well as coefficient of length of the day in its index calculations (Huglin, 1986).

4.2 OLS

According to literature (Bole, 2010; Wang et al., 2005; Tu & Xia, 2008) calculation of OLS regression is a common step during the implementation of GWR. Such regression can not only predict the dependent variable (at a global level) but it can also serve as a reference when inspecting the GWR results. In this research OLS was applied separately from GWR in order to research its predicting capabilities more thoroughly.

To get first insight, the relationships between the dependent variable and independent variables were first examined by inspecting the relationships between variables on 2 dimensional plots. OLS was conducted using SAM software (SAM, 2012), where "Multi model interference" tool was applied. The tool enables simultaneous evaluation of multiple OLS models, which derive from combinations of variables as selected by the user.

The procedure of evaluation of multiple OLS models was applied for 2007 – 2011 data. The output of the tool includes a list of models where the following results that are assigned to every model were inspected in details:

- Variables included in the model,
- R^2,
- AIC.

The derived models were assessed according to values of the criterions given above. The most explanatory models for overall yield and for Rebula variety were identified.

4.3 GWR

SAM software (SAM, 2012) was used to derive GWR as well. There are several settings that need to be set for the GWR method. These settings were eventually selected based on literature research and on testing using 2009 dataset as it includes the highest number of data records.

Spatial weighting function:
The bi-square spatial weighting function was applied as it outperformed Gaussian function and moving window function (which are available in SAM software) during the testing.

Kernel type:
An adaptive spatial kernel was applied, because this option enables the measurement of bandwidth in units of neighbouring cells instead of in units of neighbouring distance. Following this, one can specify bandwidth by ratio of neighbouring observations to all observations instead of in actual distance form regression point. According to literature (Bole, 2010) this is particularly useful if observations are not evenly distributed. In our research this is the case.

Bandwidth optimization:
An optimization using The Golden Section Search and AIC was applied. As a consequence, GWR bandwidth that was eventually applied in a model was the bandwidth which resulted in the best AICc score among all bandwidths within the specified range of its neighbours' percentages. In our case, the bandwidth for overall yield was searched for between 10 % and 75 % of regression point's closest observations. The bandwidth for Rebula variety was searched for between 10 % and 99 % of regression's point's neighbours. The maximum value for Rebula variety is very high because of the scarcity of observations. The bandwidth which resulted in the best AICc scores was eventually applied.

The GWR results consisted of statistics in tables and of plots and maps. First, AIC, R^2 and R^2 Adj were inspected and compared to the ones from OLS. A model with lower AICc scores and higher R^2 and R^2 Adj values would indicate its superiority comparing to the other one (Bole, 2010).

An examination of plots, histograms and cartographic output followed. Mean as well as local R^2 values and residuals were inspected thoroughly in order to attempt to identify the geographical areas, where variation is well or badly explained.

4.4 SOM

SOM are not a classical estimation method, therefore in this section, we first focus on why and how SOMs were applied in this study. Then, the data related issues are briefly discussed. Experimentation process and techniques of reviewing experimentation results of search for optimal SOM are presented next. Following this, the approach to project SOM to a temporal dimension and to geographic space is described. Finally, techniques to review the relation between clusters and the dependent variable of this research are presented.

SOMs were trained and inspected in the R free statistical software (R, 2012), using the "kohonen" package. Geographic projections and maps (cartography) were done using ArcGIS.

Reasons for applying SOM and general approach
As stated in the literature review (section 2.6), SOMs have been used for classification of various phenomena. Here, SOMs are used as a method of data exploration. They are applied to discover whether clusters of data records with similar characteristics can be identified from the data. The SOM data exploration method is also applied because it is rather different from the other two methods used within the study and, hence, it might provide new and complementary information. SOMs are a type of unsupervised neural network while the other two exploration methods are regressions. Finally, if the unsupervised SOM method were capable of clustering the data, a supervised SOM method could be researched in the context of this problem. Supervised SOM is capable of predicting a dependent variable's cluster according to independent variables of a data record.

The method is applied using only independent or yield explanatory variables. This is the yield per plant (Y_PER_PLA) and is not used in the training phase. The SOMs are trained using the data from the year with the most records (2009), while the data from the remaining years are mapped to a trained network. The resulting clusters are projected into geographical space in order to inspect whether the data clusters form any pattern in space. Finally, the results of the clustering are analyzed by inspecting the characteristics of a dependent variable within the identified

clusters, because this could reveal relations between clustered independent variables and the dependent variable. The results are analyzed for the whole period from 2007 to 2011.

Data related issues

The data obtained for this research includes a number of qualitative variables, which according to the literature (Vesanto, 1997), can be applied in SOM along with quantitative data. The qualitative variables obtained were however eventually not used in SOM training as it was not possible to prepare them in a form that is suitable for application in SOM. The limitations and the reasons for this decision are presented in the discussion chapter. Eventually, the same nine variables were employed for training the SOM as the ones used for the GWR and OLS regressions.

The SOM method requires scaled input data (Vesanto, 1997). Data scaling was the only preprocessing step in this study applied additionally for SOM.

Testing of SOM properties

The SOM was parameterized using the 2009 dataset because it has the highest number of data records. Consequently, this particular dataset should be the most suitable to train SOM, as more data records should result in more accurate clustering.

The testing of SOM parameters was based on the parameters found in literature (Kohonen (1990); Vesanto et al. (2000); Cereghino et al. (2005); (Wehrens and Buydens, 2007)). A short presentation of possible parameters is given below.

Dimension of the map:

According to Vesanto et al. (2000) the suitable size of SOM is 5 * sqrt (n) where n is the number of training samples. In 2009 data there are 1225 training samples. Therefore approximately 175 neurons would be suitable. The number of training samples for Rebula variety in 2009 data is 136, therefore 58 neurons would be suitable. In both cases the suitable shape would be rectangular in order to orientate from results. A number of dimensions were tested for SOM for all grape varieties and for SOM for the Rebula variety, because the quality of clustering is dependent of the appropriateness of the map dimension.

Learning rate:

According to Kohonen (1990), alpha should start with a value that is close to unity and then decrease to values of 0.01 or less over a long period. On the other hand, the preset value for alpha in Matlab software is 0.5 (Vesanto et al., 2000), while the preset alpha value in Kohonen package in SOM is 0.05 (Wehrens and Buydens, 2007). Various beginning alpha values were therefore tested. Alpha value at the end of the training was always set to 0.01. The decline of alpha value during the iterations was always set to linear.

Number of iterations:

According to literature (Kohonen, 1990), the number of optimal iterations is dependent of the number of neurons in the output map and should be at least 500 x number of neurons. Various numbers of iterations were tested nevertheless.

Radius of the neighbourhood:

The starting radius of the neighbourhood was set to cover 2/3 of all distances at the beginning, as Kohonen (1990) points out that the initial radius can be more than a half of the diameter of the network. The neighbourhood was set to gradually decrease to one unit. This property was always set the same during the testing.

Shape of the neighbourhood:
The selected shape of the neighbourhood was hexagonal. This property was always set the same during the testing.

Visual interpretation of the derived plots was used for validation and assessment of the experimentation results. The most information bearing SOM output was the U-matrix. This matrix shows the sum of the distances of a particular neuron to all its immediate neighbours. This information was used to identify the clusters. Eventually, the SOM which resulted in U-matrix with the clearest clusters was chosen for further research. Component planes were analyzed by comparing them to U-matrix in order to review the distribution of vector weights on the network and to assess the influence of individual variables on clustering.

Furthermore, the number of occurrences of data records within neurons was inspected to observe the distribution of data records assigned to individual neurons. Mapping quality of individual neurons or mapping areas was estimated by observing the plot which shows a mean distance of objects mapped to a unit to the codebook vector of that unit. A good mapping should show small distances everywhere in the map. Finally, the plot of mean distance to the closest codebook vector during the training was examined in order to asses the progress of training and suitability of number of iterations.

SOM clustering and its projection to geographic space and to temporal dimension
Based on testing results, one SOM was trained for entire data (1225 records) and one SOM was trained for Rebula variety data (136 records) of 2009. Clusters were derived from SOMs by visual inspection of U-matrices. The neurons with the highest sum of distances to its immediate neighbours were regarded as the neurons which lie on a border of a cluster. Following this approach, the SOM was delimited to a number of clusters and finally, the data records were given the label of the clusters where they fall.

Clusters were projected into a geographical space by displaying the vineyard centroids, coloured according to clustering. The derived map was visually inspected to asses whether there is any relation between clustering and geographical location of vineyards.

SOM training was not applied for each year separately. Instead, the data from 2007 – 2011 was mapped to a trained network trained using 2009 data. For example, each data record of 2007 data was assigned to that neuron of a trained network, which was described by a vector with weights similar to the weights of the vector of that particular 2007 data record. This technique therefore attributed the data records of 2007 – 2011 with cluster number based on SOM trained from the 2009 data.

The relation between the derived clusters and yield
The relation between the derived clusters and the dependent variable was researched by calculating basic descriptive statistics of data records within clusters. The main goal was to asses whether the data clustering reflects the variation of the dependent variable.

Mean yield per plant per cluster was compared particularly to the clusters of the same year in order to assess whether there is a relation between yield and the cluster to which the data were assigned to. The range of yield per plant within clusters and standard deviation of yield within a cluster was reviewed in order to supplement the findings that derived from the comparison of mean yield per plant per clusters.

4.5 Comparison of OLS and GWR results with meteorological characteristics

Meteorological data were initially acquired in order to provide explanatory variables for this research. Because this proved to be impossible, a different examination of the relation of meteorological data to yield quantity estimation was applied.

Meteorological data were used to estimate its influence on yield estimation accuracy on a yearly basis. More precisely, they were used to identify whether yearly (seasonal) meteorological characteristics lead to better or worse prediction capabilities of a model. The comparison was made by first, calculating Pearson's correlation matrix using selected meteorological data and R^2 values of the most explanatory OLS and GWR models and second, by visually inspecting it. The data used for the research were mean temperatures, precipitation quantity and number of sun hours, from April to including September from the Bilje meteorological station. The regression results of both all yield and Rebula variety were applied.

5. RESULTS AND DISCUSSION

This chapter follows the same order than the methodology chapter. That is, data preprocessing results are described first and they are followed by OLS, GWR and SOM results. The result of comparing GWR and OLS results to meteorological data follows. Each section contains the results for the entire yield and for the Rebula variety. Individual results are discussed in each section and a general discussion concludes this chapter. The source of data in figures and tables is MAE, if not cited differently

5.1 Data preprocessing

The most important data preprocessing steps as well as intermediate and final preprocessing results are presented and discussed because they considerably influence the size and the content of the data used in the research. Meteorological data preparations results however are not presented as they do not influence the dataset applied in the research as much. Nevertheless, meteorological data preparation problems are mentioned in the discussion that follows because it is important to explain why meteorological data were not applied in greater extent in this research.

Data joining

The joining of datasets resulted in a considerable decrease of data records that were applicable for this research. A decrease is depicted in Figure 19 and Figure 20, where the number of vineyard polygons, number of vineyard characteristics records and number of yield declaration data records between 2007 and 2011 are shown. A comparison can be made with another category shown in these figures, which is the final number of data records that were actually applied. The decrease as seen in the figures occurred mainly due to joining of the datasets, because the remaining data preparation tasks further decreased the number of data records for only few percents.

In Figure 19 one can see that there were approximately 3000 vineyard polygons in the acquired data for every year. As there are approximately 8000 vine characteristics data records per year, one can conclude that there are between two and three vine characteristics records per vineyard on average. One can also observe that there were on average approximately 4500 yield declarations per year. Regardless of a few thousand data records in each category, there were only approximately 1160 data records per year on average applied for our research, after the data preprocessing phase was concluded.

Slight trends in increase of vineyard polygons and in increase of vineyard characteristics records in the data can be observed. Number of yield declaration records and number of applicable data records on the other hand show a different trend, as in both cases the highest number of records was in 2009 and has been decreasing since.

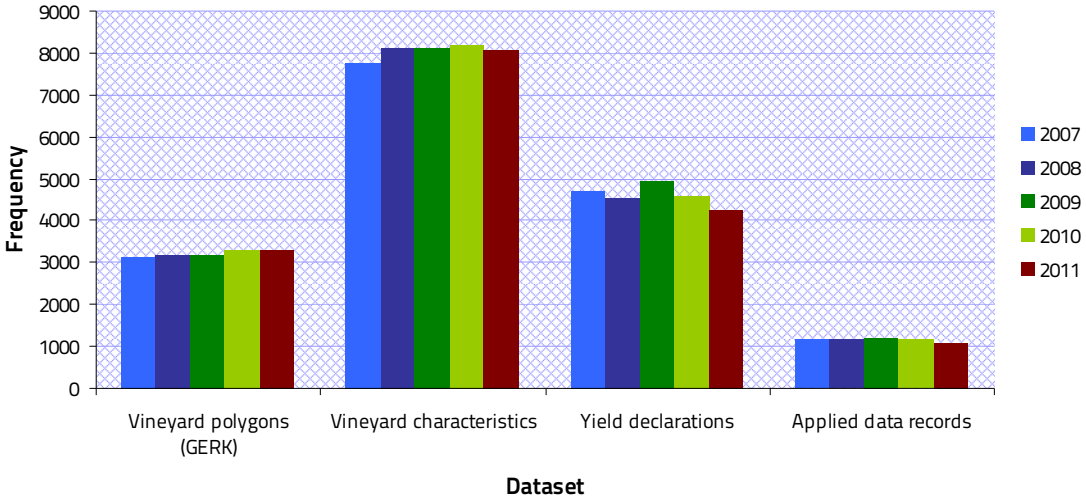


Figure 19. Comparison of acquired and applied data records frequency (2007 – 2011 data).

Figure 20 shows the number of acquired and applied data records of Rebula variety. There were approximately 1300 vineyard polygons and approximately 1850 vineyard characteristics records available in the acquired data. Furthermore, the number of yield declaration records was usually bellow or slightly over 1000. However, there were only between circa 120 and 130 data records applicable for research after preprocessing was concluded.

There appears to be no trend visible in the number of vineyard polygons and number of vineyard characteristics records when observing their change. On the other hand there appears to be a negative trend in a number of yield declaration records and in number of applicable data records, though the maximum number in both categories was in 2009.

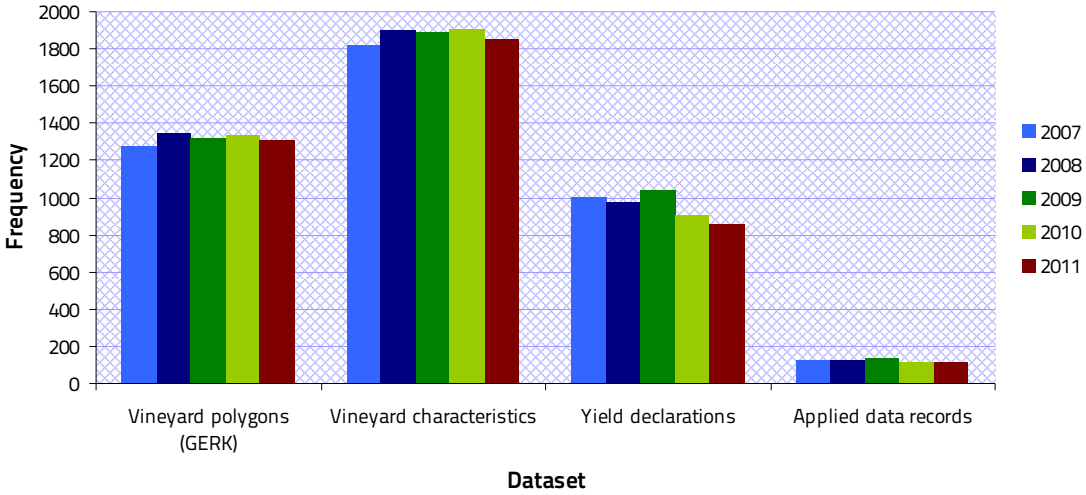


Figure 20. Comparison of acquired and applied data records frequency (2007 – 2011 Rebula data).

Data manipulation

The data manipulation phase included various tasks out of which only the ones that influenced the size of data the most are mentioned here. The removal of outliers resulted in a decreased number of data records. The decrease was minor, comparing to the decrease during data joining. Nevertheless, approximately 5 % of data records were removed for every year, taking into account their removal based on a dependent variable (YI_PER_PLA) and explanatory variables (Z_AVG,

R_DIST, Y_PLANTING) outliers. The criterion for exclusion of outliers for YI_PER_PLA and Y_PLANTING are mentioned in methodology (section 4.1), while outliers in Z_AVG and R_DIST were searched for using histograms as the one depicted in Figure 21. One can see the outlying data records on both ends of histogram. The data records with row distance lower than 2 metres and higher than 3,3 metres were removed in this particular case.

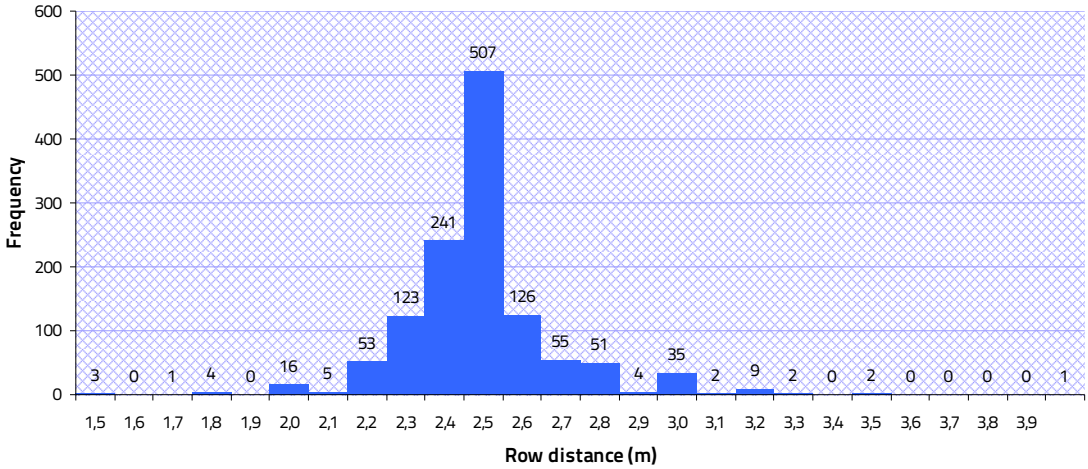


Figure 21. Histogram used for detection of outliers in row distance (R_DIST) variable (2009 data).

The data records that were left after the removal of the outliers were eventually applied in the research. Because we are dealing with estimating yield per plant, the information about the ratio of plants used in the research to all plants is a good indicator of the amount of explanatory data eventually used in the research. Details about this category in the period between 2007 and 2011 can be observed in Table 4 and Table 5.

In Table 4, the ratio of plants applicable for research to all plants is presented in bold, as it is the most important information. As one can observe, the ratio is between 15 and 17 %, which is low. The table also shows that the number of plants in acquired data is increasing steadily. However, the ratio of applicable plants as well as the number of applicable data records is, with exception of 2009, decreasing. In general the table shows that there appears to be a trend in the decrease of suitability of the data (the number of data records that can be applied) for the purpose of our research.

Table 4. Information about quantity of applied data (2007 – 2011).

	2007	2008	Index 07/08	2009	Index 08/09	2010	Index 09/10	2011	Index 10/11
Plants all	6041576	6175547	102.22	6270653	101.54	6324906	100.87	6346108	**100.34**
Plants applied	1013469	1024608	101.10	1069245	104.36	1013356	94.77	959515	94.69
Plants applied %	**16.77**	**16.59**	**98.91**	**17.05**	**102.77**	**16.02**	**93.96**	**15.12**	**94.37**
# applied data records	1169	1154	98.72	1225	106.15	1152	94.04	1084	94.10

Table 5 shows the same information for Rebula variety data. The ratio of data records that can be applied in the study is even lower and with 7.3 – 8.8 % hardly represents the data. The number of Rebula plants in the study area is in general increasing, though the trend has stalled in the last

two years. However, the ratio of plants that are applicable for our research and the number of the data records that can be used in the research, are decreasing. The exception though is 2009.

Table 5. Information about quantity of applied data (2007 – 2011, Rebula).

	2007	2008	Index 07/08	2009	Index 08/09	2010	Index 09/10	2011	Index 10/11
Plants all	1271581	1300836	102.30	1340269	103.03	1332535	99.42	1339710	100.54
Plants applied	112436	112851	100,37	116916	103.60	102976	88.08	98397	95.55
Plants applied %	**8.84**	**8.68**	**98.11**	**8.72**	**100.55**	**7.73**	**88.59**	**7.34**	**95.04**
# applied data records	129	128	99.22	136	106.25	122	89.71	117	95.90

Multicollinearity checks using VIF measure of collinearity indicated the presence of collinearity between two pairs of variables. This collinearity can be observed from Pearson's correlations matrices as well. In Table 6 and Table 7, where the correlation values are shown, the correlation between variables is coloured (green – weak, yellow – medium, red - strong) for a clearer presentation. Matrices derive from the 2009 data.

Table 6 indicates that in 2009 only V_SPACING has certain correlation with yield per plant, though this correlation is weak. The correlation of other explanatory variables with dependent variable practically does not exist. One can however observe that there is a strong correlation of explanatory variables PLANT-SUM with NET-AREA and F-AREA-SUM with F-PLANT-SU. There are also certain weak correlations between explanatory variables, for example SLOPE_AVG and Z_AVG and between Y_PLANTING and V_SPACING.

Table 6. Pearson's correlation matrix of MAE data and its derivatives (2009 data).

	YI_PER_PLA	R_SPACING	V_SPACING	Y_PLANTING	Z_AVG	SLOPE_AVG	EXP_AVG	PLANT-SUM	PARC-AREA	NET-AREA	F-AREA-SUM	F-PLANT-SU
YI_PER_PLA	**1**	**0.155**	**0.311**	**-0.201**	**-0.083**	**-0.055**	**-0.041**	**-0.155**	**-0.011**	**-0.05**	**-0.065**	**-0.092**
R_SPACING	0.155	1	0.249	-0.153	-0.132	0.002	-0.046	0.035	0.119	0.217	0.141	0.095
V_SPACING	0.311	0.249	1	-0.589	0.044	0.079	0.055	-0.248	-0.026	0.054	-0.147	-0.211
Y_PLANTING	-0.201	-0.153	-0.589	1	-0.037	-0.024	-0.057	0.225	0.126	0.04	0.245	0.279
Z_AVG	-0.083	-0.132	0.044	-0.037	1	0.494	0.283	-0.205	-0.209	-0.208	-0.204	-0.211
SLOPE_AVG	-0.055	0.002	0.079	-0.024	0.494	1	0.401	-0.176	0.055	-0.156	-0.012	-0.04
EXP_AVG	-0.041	-0.046	0.055	-0.057	0.283	0.401	1	-0.088	0.117	-0.078	0.023	0.008
PLANT-SUM	-0.155	0.035	-0.248	0.225	-0.205	-0.176	-0.088	1	0.285	0.947	0.358	0.382
PARC-AREA	-0.011	0.119	-0.026	0.126	-0.209	0.055	0.117	0.285	1	0.294	0.552	0.535
NET-AREA	-0.05	0.217	0.054	0.04	-0.208	-0.156	-0.078	0.947	0.294	1	0.334	0.335
F-AREA-SUM	-0.065	0.141	-0.147	0.245	-0.204	-0.012	0.023	0.358	0.552	0.334	1	0.982
F-PLANT-SU	-0.092	0.095	-0.211	0.279	-0.211	-0.04	0.008	0.382	0.535	0.335	0.982	1

In Table 7 one can observe the Pearson's correlation matrix for Rebula variety for 2009. The comparison of this correlation matrix to the one in Table 6 shows that there are even weaker correlations between the dependent variable and explanatory variables. In the case of Rebula variety, the dependent variable is correlated the most highly with SLOPE_AVG. The correlations between the independent variables are slightly stronger in case of Rebula variety than in the case of the whole yield. Similarly as for the whole yield, there are very high correlations between first PLANT-SUM & NET-AREA and second, between F-AREA-SUM & F-PLANT-SU.

Table 7. Pearson's correlation matrix of MAE data and its derivatives (2009 Rebula data).

	YI_PER_PLA	R_SPACING	V_SPACING	Y_PLANTING	Z_AVG	SLOPE_AVG	EXP_AVG	PLANT-SUM	PARC-AREA	NET-AREA	F-AREA-SUM	F-PLANT-SU
YI_PER_PLA	**1**	**-0.071**	**0.046**	**0.094**	**-0.067**	**-0.171**	**-0.027**	**-0.137**	**0.093**	**-0.099**	**0.079**	**0.068**
R_SPACING	-0.071	1	-0.027	-0.581	-0.107	0.04	0.007	0.134	0.193	-0.021	0.347	0.383
V_SPACING	0.046	-0.027	1	0.091	-0.211	0.086	-0.077	0.148	0.195	0.284	0.258	0.211
Y_PLANTING	0.094	-0.581	0.091	1	0.126	0.091	0.129	-0.144	-0.005	0.12	-0.294	-0.35
Z_AVG	-0.067	-0.107	-0.211	0.126	1	0.3	0.044	-0.087	-0.355	-0.074	-0.456	-0.441
SLOPE_AVG	-0.171	0.04	0.086	0.091	0.3	1	-0.071	<.001	-0.138	0.032	-0.114	-0.14
EXP_AVG	-0.027	0.007	-0.077	0.129	0.044	-0.071	1	0.044	0.143	0.061	-0.007	-0.027
PLANT-SUM	-0.137	0.134	0.148	-0.144	-0.087	<.001	0.044	1	0.402	0.958	0.347	0.354
PARC-AREA	0.093	0.193	0.195	-0.005	-0.355	-0.138	0.143	0.402	1	0.411	0.44	0.413
NET-AREA	-0.099	-0.021	0.284	0.12	-0.074	0.032	0.061	0.958	0.411	1	0.291	0.278
F-AREA-SUM	0.079	0.347	0.258	-0.294	-0.456	-0.114	-0.007	0.347	0.44	0.291	1	0.979
F-PLANT-SU	0.068	0.383	0.211	-0.35	-0.441	-0.14	-0.027	0.354	0.413	0.278	0.979	1

Out of the pairs of variables identified as multicollinear (PLANT_SUM and NET_AREA; F_AREA_SUM and F_PARC_SU), the variable with the worse correlation to the dependent variable was abandoned. The whole period of 2007 – 2011 was taken into account. The following variables were eventually used in the research:

- spacing between rows (R_SPACING),
- spacing between vines (V_SPACING),
- year of planting (Y_PLANTING),
- average altitude (Z_AVG),
- average slope (SLOPE_AVG),
- average exposition (EXP_AVG),
- vineyards area (PARC-AREA),
- vineyards area covered with vine in question (NET-AREA),
- vineyards area on farm (F-AREA-SUM).

Discussion

Two facts that are likely to influence the results of the research are obvious after reviewing the data preprocessing results. First, the ratio of the data that can be used for the research is low especially for Rebula variety. Second, the correlations between the dependent variable and explanatory variables are very weak, one could even claim that they are random.

Only up to 15 % of data records can be used for the research, in case of Rebula variety the ratio is even lower at 7%. In both cases the ratio of plants whose characteristics can be used for estimation of yield is only slightly higher. In the case of Rebula variety however, the number of records available for the research is even more troublesome than the ratio. One can question whether the size of just over 100 data samples can provide an accurate modelling result. One can thus conclude that the ratio of usable data and in case of Rebula also the number of data records available for research are a limiting factor for accuracy of estimation. Furthermore, the fact that the ratio of applicable data records is in a decreasing trend is troublesome as well. If this continues, the estimation of yield per plant could become even harder in the future.

According to correlations between the dependent variable and explanatory variables, the estimation of yield per plant using the set of variables used in this research is a challenge even now. On the other hand, the data mining by its definition is a method which can find patterns even in cases where there appear to be none. Furthermore, Pearson's correlation matrix shows global correlations between the variables, thus local regression methods, such as GWR might find patterns in data on a local scale.

According to literature, the meteorological data could provide additional explanatory variables that are important in the explanation of yield variation. For that purpose, meteorological data from various meteorological stations were acquired. Its accurate interpolation to such a small spatial unit as a vineyard, however does present a problem. According to the literature (Belec et al., 1998), the temperatures in most of Slovenia decrease with a certain vertical gradient while the rainfall increases with the vertical gradient. It was therefore expected that such patterns could be found in acquired meteorological data. Consequently the vineyards data records located on known locations could be attributed with meteorological variables.

However, after analyzing the data, the assessment was made that it is not possible to calculate a reliable interpolation of meteorological characteristics based on altitude. To come to such a conclusion, the relations between monthly mean temperatures and altitude and between the monthly sum of precipitation and altitude were examined. The relation between these variables using SEA and PARS meteorological data can be seen in Figure 22 and Figure 23 below. The figures are based on 2009 data, because it was the most complete.

If one observes Figure 22, one can see that the temperature does not decrease with altitude in a clear trend. Another issue that can be observed from the figure is the difference between SEA and PARS measurements. Though Bilje (55m) and Vipolže (60 m) are approximately 11 km away and both located in flat area, their difference in mean temperature is between 1 and 2 °C in all months. It actually varies even more than between Vipolže (60 m) and Višnjevik (220 m) stations, between which there is more than 150 meters difference in altitude. Further, even within PARS data, there appears to be only partial correlation between temperatures and altitude. For example, the highest temperatures are in most cases measured in Vipolže at 60 meters altitude, while the lowest temperatures in general are measured in station Šlovrenc with 100 meters altitude. Moreover, in August for example, the mean temperatures measured in all PARS stations are practically the same.

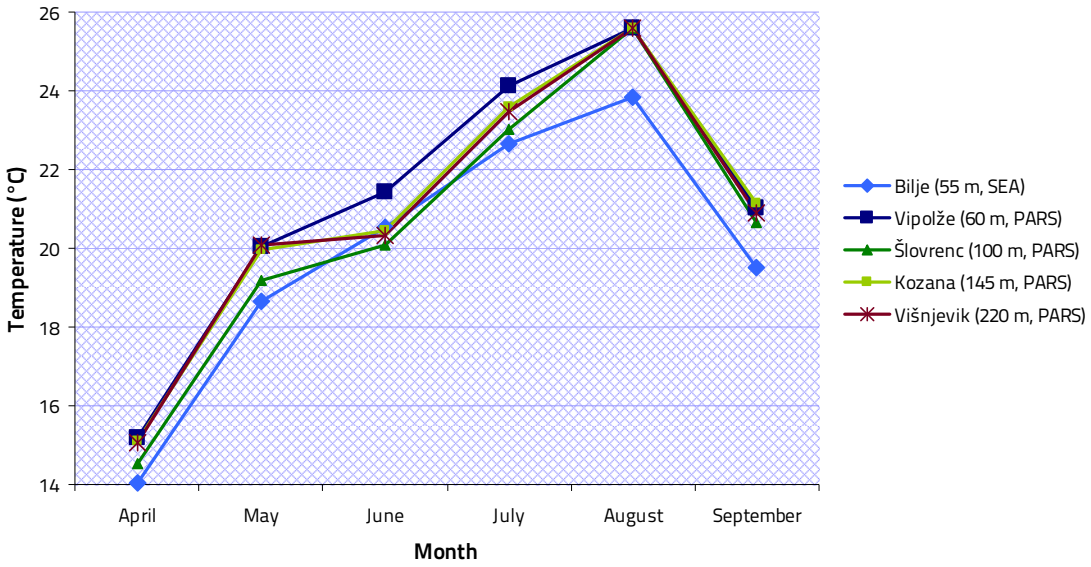


Figure 22. Relationship between temperature and altitude in 2009. Source: SEA (2012), PARS (2012)

Figure 23 shows that the relationship between altitude and precipitation is stronger than the one between temperature and altitude. In most months the precipitation quantity in most cases indeed roughly increases with altitude. However in none of the months does the order of meteorological stations stay the same if one orders them by altitude and by precipitation. Further, in August, when the droughts are the most common in the study area, and therefore the precipitation quantity is the most important, the order differs the most.

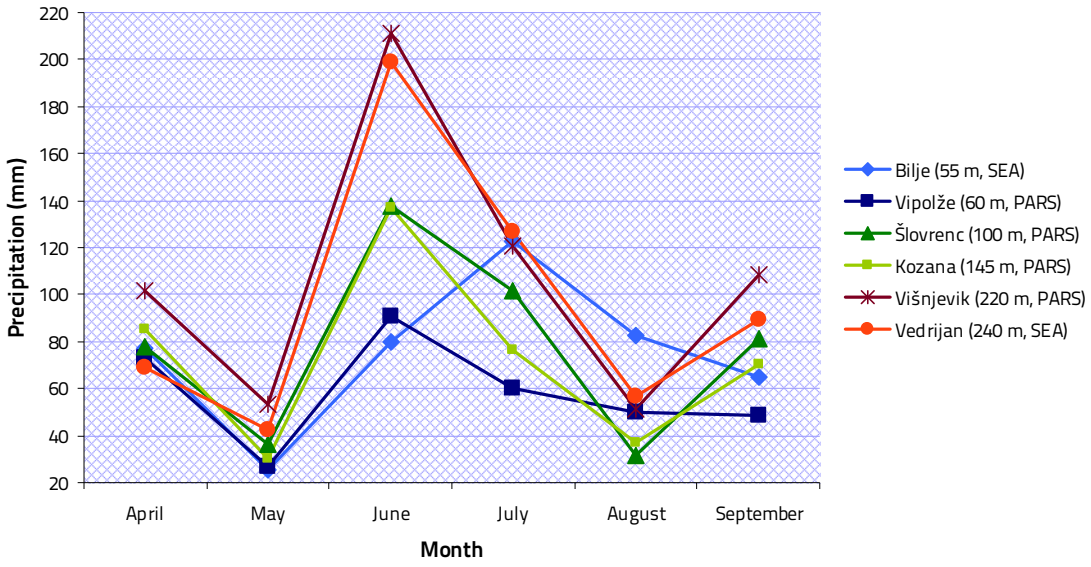


Figure 23. Relationship between precipitation and altitude in 2009. Source: SEA (2012), PARS (2012)

The goal of acquisition of this data was to estimate meteorological characteristics on very small scales, that is at the level of vineyards. Though a certain correlation between precipitation and temperature on the one hand and altitude on the other hand altitude is evident, the data presented above can not be applied in this research as a yield explanatory variable because these phenomena can not be accurately projected onto the scale of vineyards.

5.2 OLS

OLS is a global regression method, therefore the plots presented in Appendix A, showing correlations between all possible pairs of a dependent variable and explanatory variables, can present an introductory insight into its prediction power. The plots depict the correlations for all data and for Rebula variety for 2009, using data after preprocessing. They are a graphical representation of the data presented in Pearson's correlation matrices in Table 6 and Table 7. By observing the plots one can see that none of the explanatory variables have a linear or any other kind of relationship to a dependent variable. All of the relationships can be described as 'noisy".

If one plots a regression line onto the plots in Appendix A (red line in plots) some conclusions can be made. Firstly, some relationships are in fact positive; some are negative (though the relationships appear to be random). For example, the yield per plant increases with row spacing and vine spacing, while it decreases with the sum of plants (with the same characteristics) in the vineyard. Secondly, there is a difference between the relationships for the entire yield and for Rebula variety. For example, the above mentioned row spacing and vine spacing does not have as significant influence on the yield of Rebula variety as it does on the yield in general. Furthermore, vineyards area on farm is negatively correlated with the yield per plant of the entire yield, while it is positively correlated with the yield per plant of Rebula variety.

OLS results are presented in Table 8, Table 9 and in Figure 24. They derive from the comparison of all possible combinations of explanatory variables for each dataset, which results in 511 models for each dataset. For example, among 511 possible models for explanation of Rebula yield with 9 explanatory variables for 2011, only the two models with the best results are presented below.

All data

Table 8 presents the comparison of the best OLS models for 2007 – 2011 data according to R^2 and AIC criterions. This table shows that R^2 values are very low in all cases. Following this, none of the models for none of the years explain enough variation of the dependent variable to be assessed as feasible. One can also observe that the dependent variable can be explained with almost the same accuracy even when some explanatory variables are not used in a model. This conclusion can be made by comparing R^2 values of the same year.

Table 8: Review of OLS yield estimation capability (2007 – 2011).

	2007 best R^2	2007 best AICc	2008 best R^2	2008 best AICc	2009 best R^2	2009 best AICc	2010 best R^2	2010 best AICc	2011 best R^2	2011 best AICc
R^2	0.1	0.099	0.082	0.08	0.125	0.124	0.141	0.14	0.109	0.106
AICc	13943	13940	13760	13753	14594	14587	13604	13600	12750	12743
# of variables in a model	9	7	9	5	9	5	9	6	9	4

Rebula variety data

In Table 9 one can compare the results of the OLS analysis for Rebula variety. None of the models can explain the dependent variable with a desired level of accuracy for none of the years. One can also observe that in 2008 and 2009, the best AICc score was scored by a model with only one dependent variable, though that variable practically does not explain the variation of the dependent variable at all.

Table 9. Review of OLS yield estimation capability (2007 – 2011, Rebula).

	2007 best R^2	2007 best AICc	2008 best R^2	2008 best AICc	2009 best R^2	2009 best AICc	2010 best R^2	2010 best AICc	2011 best R^2	2011 best AICc
R^2	0.103	0.071	0.07	0.037	0.09	0.029	0.135	0.09	0.09	0.051
AICc	1536	1527	1570	1556	1625	1616	1427	1420	1364	1353
# of variables in a model	9	3	9	1	9	1	9	3	9	2

Discussion

The results of the OLS revealed that global regression can not explain the variation in yield, at least not using the explanatory variables used in this research. Poor results were expected after the correlation matrices (Table 6 and Table 7) and correlation plots (Appendix A) were analyzed. Low correlations between the dependent variable and the explanatory variables raise the question of suitability of most explanatory variables for research dealing with modelling yield per plant using global regression.

One can also conclude that poor results are not due to the fact that a type of grape variety was not considered in the analysis as the explanatory variable, as the results are similarly poor (in fact even worse) when OLS is applied using data for only one grape variety (Figure 24). This could on the other hand be a consequence of a small number of data records used for prediction of Rebula yield.

A comparison between the results of entire yield and Rebula yield OLS regression is depicted in Figure 24 where R^2 of the most explanatory model for entire yield and for Rebula variety are compared. One can see that with the exception of 2007 the results are a bit better for entire yield than for Rebula variety. However, R^2 in all cases is very low.

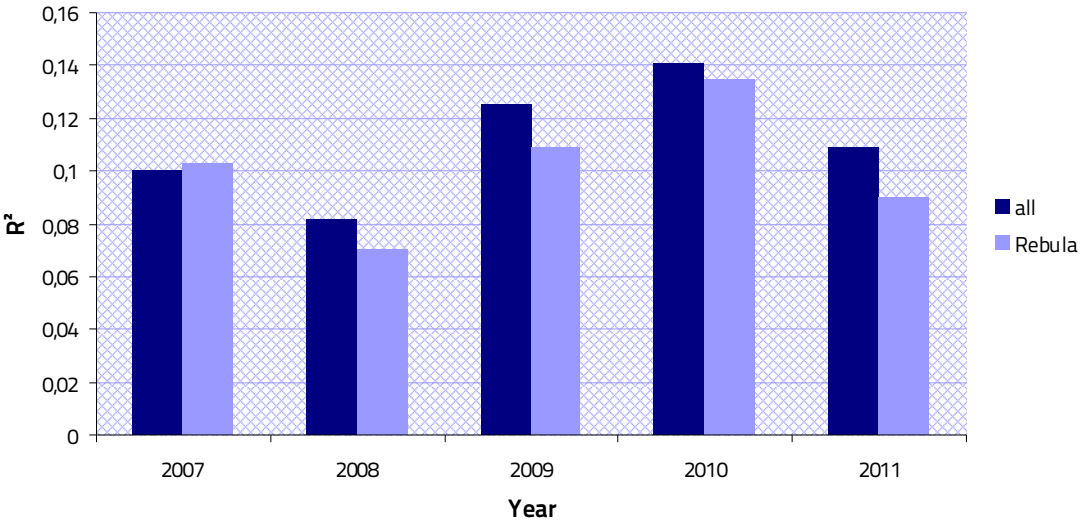


Figure 24. Comparison of R^2 of the best OLS models (2007 - 2011).

5.3 GWR

Two main types of GWR results were obtained from the analysis. First, there are statistics and plots, which describe the result with the intention of summarizing the model. These are compared to the results of OLS method. Second, there are maps, which enable an insight into the geographical component of the model. Both types of results are presented and discussed for all data and for the Rebula variety data.

All data

Table 10 presents the results of GWR and OLS for all data for 2007 – 2011. The first row shows the number of data records (locations) in each year. One can see that this number varies and that approximately 1160 samples is a mean number of samples over the years. The second row presents the results of AICc, which differ from year to year. What is more important though is that AICc of GWR and OLS models differ for the same years. The difference of approximately 25 – 60 scores in favour of GWR (lower AICc) means that the GWR model is considerably superior to OLS for all years. The third row shows R^2 values, which vary from year to year as well and again are considerably in favour of GWR for all years. However, even in 2010, when the R^2 value for GWR is the highest (0.283), it is still too low to asses the GWR model as feasible. R^2 Adj results are a bit lower than for R^2, but show the same pattern as R^2 results. The last row shows the percentage of neighbours of each regression point of a model, which were used to predict its value. As OLS is a global regression method, all data records were considered when predicting the value of dependent variables. For GWR however, the percentage of neighbours used varies from 27 to 67 percent. If one compares the percentage of neighbours used to R^2 value, one can see that three of the best models with R^2 0.24 or higher (2007, 2009 and 2010), were constructed using approximately 30 % of neighbours. The remaining two models which, resulted in worse R^2 values (2008 and 2011), were constructed using approximately twice as many neighbours.

Table 10. Comparison of GWR and OLS models' diagnostic statistics (2007 – 2011).

	GWR 2007	OLS 2007	GWR 2008	OLS 2008	GWR 2009	OLS 2009	GWR 2010	OLS 2010	GWR 2011	OLS 2011
# of Locations	1169	1169	1154	1154	1225	1225	1152	1152	1084	1084
AICc	13911	13943	13736	13760	14559	14594	13547	13604	12715	12750
R^2	**0.263**	0.1	**0.143**	0.082	**0.24**	0.125	**0.283**	0.141	**0.189**	0.109
R^2 Adj	0.199	0.094	0.117	0.075	0.195	0.119	0.234	0.135	0.159	0.102
% of neighbours used	27.2	100	67.4	100	36.4	100	33.8	100	60.6	100

In Figure 25 one can see descriptive statistics of actual and GWR estimated yield per plant for 2007 – 2011. One can see that the estimated minimal values are considerably higher than the actual ones and that estimated maximum values are considerably lower than the actual ones. Estimated mean values on the other hand are approximately the same as the actual ones. These ratios are reflected in standard deviation, as actual standard deviation is more than twice as high as the estimated one. Finally, models with higher R^2 (2007, 2009 and 2010) have estimated values closer to the actual ones for minimum, maximum and standard deviation. Based on the figure below, one can conclude that in our case the GWR model fails to predict extreme values.

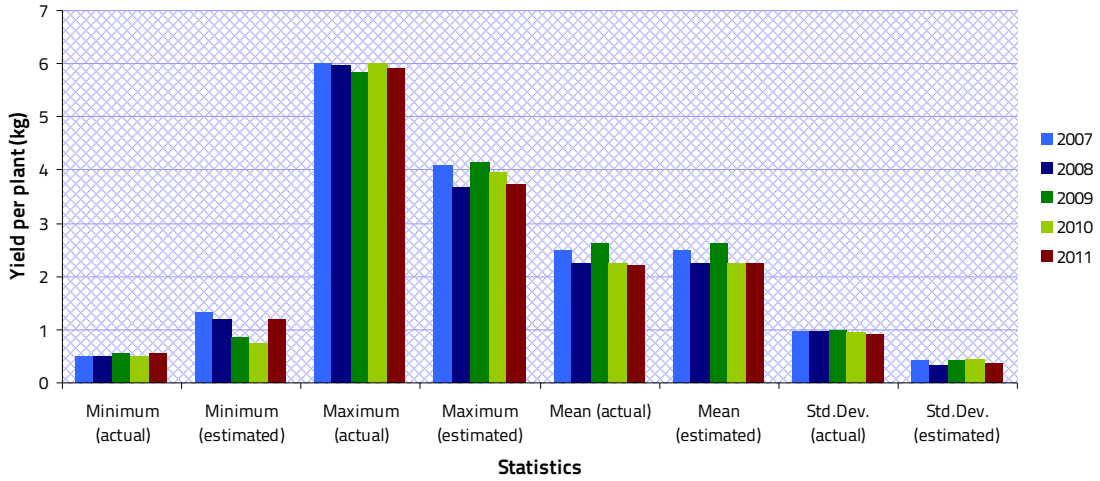


Figure 25. Statistical comparison of actual and GWR estimated yield (2007 – 2011).

Another GWR result that is particularly useful is the local R^2 which can be plotted on a map. Figure 26 shows local R^2 values for the 2007 - 2011 data. When observing the maps, one must be careful when reading the legends, as the values differ between the maps in order to better illustrate the variation of local R^2 within each year. A clear pattern in local R^2 distribution is visible in maps for all years. However, the patterns on maps showing local R^2 for 2008 – 2011 are similar, but the pattern for 2007 differs. For 2007 data, the highest local R^2 values are in the northernmost part of the area. The lowest local R^2 values in 2007 are in the southernmost, central-eastern and western part of the area. The distribution of local R^2 for 2008 – 2011 on the other hand has in common the highest local R^2 values on the southernmost and south-western part of the area. The lowest local R^2 values for 2008 – 2011 are usually in the western and eastern part of the area. However, by inspecting the legends of the maps, one can see that even the highest local R^2 values in all maps are relatively low, even in the areas where R^2 values are the highest. Following, it is not feasible to predict yield per plant in any part of the study area using this model.

Some additional GWR visualizations, which enable more detailed insight into GWR results, are shown in Appendix B. For every year from 2007 to 2011, the following graphics are presented:

1. plot of GWR estimation on actual yield per plant,
2. plot of residuals to estimated values,
3. local R^2 histogram,
4. residuals histogram,
5. map of residuals.

When one observes the visualizations in Appendix B, one can conclude that for years 2007 – 2011 in general:

1. there is a correlation between estimated and actual values of yield per plant, however there is a lot of variance in estimations and there is an unsatisfactory estimation of extreme values,
2. there is no pattern visible when plotting residuals to estimated values as there appears to be a random noise shown on all of the plots,
3. there is no pattern in local R^2 histograms when comparing them over the years as each histogram has its specific characteristics already reflected in summary statistics of the model and its local R^2 representation,

4. the residuals distribution shown in residual histograms appears to be normal, but the histograms are relatively flat (taking into account their range), meaning that the accuracy of the model is not high,
5. there appears to be no pattern on a map of residuals, as few of extremely high and extremely low, but also medium residuals appear to be randomly scattered over the area, though the effect of very high maximum values of residuals in 2008 can be seen.

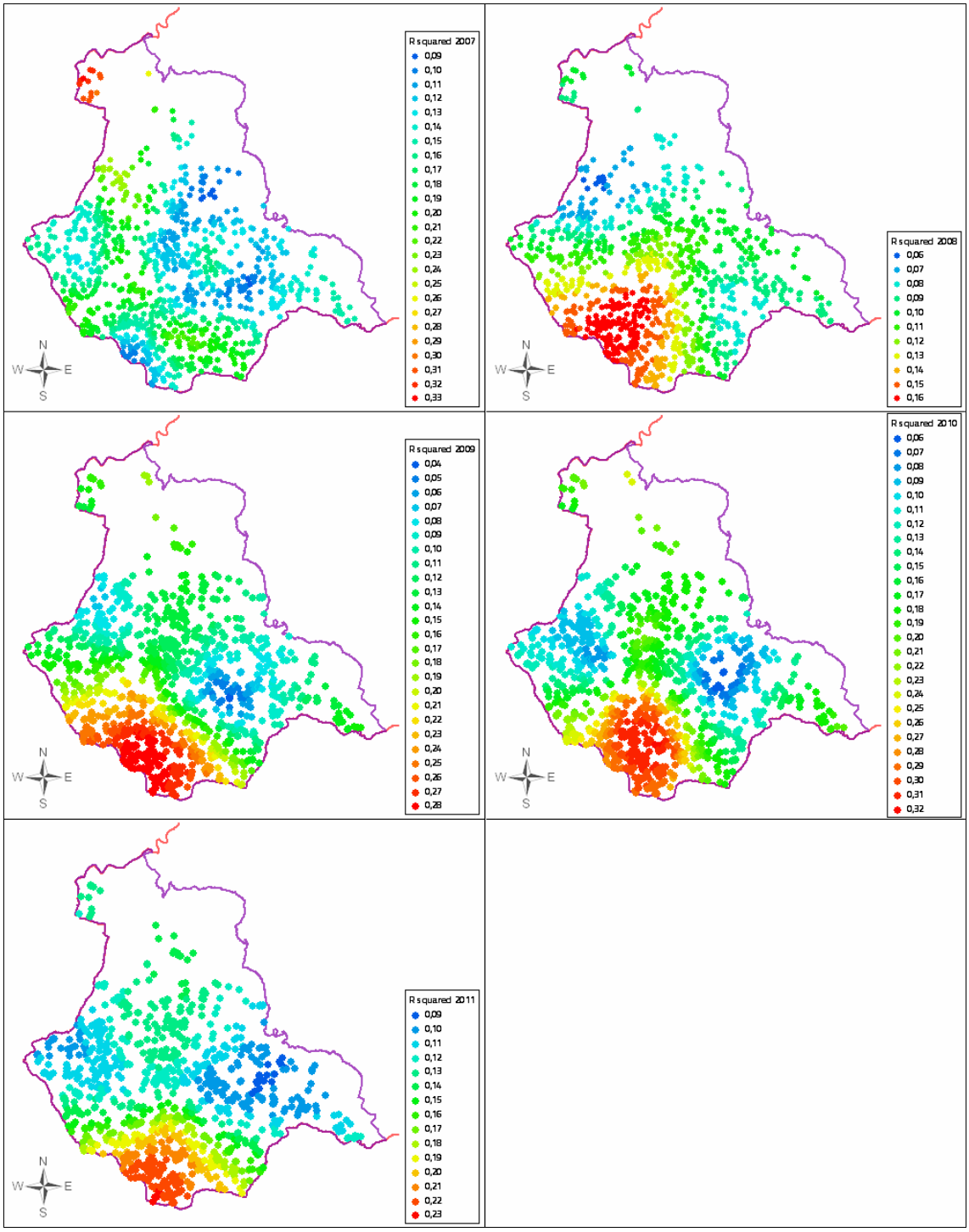


Figure 26. Local R^2 values of GWR (2007 – 2011).

Rebula variety data

The results of GWR and OLS analysis for Rebula variety are shown in Table 11. One can first see that there are 117 to 136 records available for each year. Also, the comparison of AICc values between GWR and OLS models shows that OLS model is superior to the GWR model, if the model's explanatory capabilities and its complexity are both taken into account. However, the differences between AICc scores are for example quite low for 2007 and 2011. R^2 values presented in the third row of the table indicate, that even though AICc favours OLS, the GWR model explains higher ratio of variance of the dependent variable. Furthermore, the difference of R^2 between GWR and OLS is in most years considerable and is more than twice as high in three out of five years (2007, 2008, and 2011). However, the highest R^2 values are still too low to regard a GWR (and OLS) model as feasible. For example the highest R^2 value is 0.231 for 2011, while R^2 values for the other years are somewhere between 0.15 and 0.22. R^2 Adj values are very low as well. Contrary to AICc however, R^2 Adj favours GWR, even though both indicators assess a model based on its accuracy and complexity. Finally, the last row in the table includes the percentage of neighbours used to obtain a model. In case of GWR, the maximum of the available percentage was applied, that is 99 %. This means that the technique must use all of the available data records to obtain the best GWR result based on the AICc criterion.

Table 11. Comparison of GWR and OLS models' diagnostic statistics (2007 – 2011, Rebula).

	GWR 2007	OLS 2007	GWR 2008	OLS 2008	GWR 2009	OLS 2009	GWR 2010	OLS 2010	GWR 2011	OLS 2011
# of Locations	129	129	128	128	136	136	122	122	117	117
AICc	1537.5	1535.9	1575.3	1569.4	1633.1	1625.3	1436.1	1427.3	1363.9	1363.8
R^2	**0.219**	0.103	**0.158**	0.07	**0.146**	0.09	**0.183**	0.135	**0.231**	0.09
R^2 Adj	0.12	0.043	0.051	0.007	0.044	0.032	0.076	0.074	0.121	0.023
% of neighbours used	99	100	99	100	99	100	99	100	99	100

In Figure 27 one can see the descriptive statistics of actual and GWR estimated yield per plant for Rebula variety. The estimated minimum values are more than twice as high as the actual ones. Furthermore, the estimated maximum values are considerably lower than the actual ones. Nonetheless, the estimated mean values are similar to the actual ones. The differences in actual and estimated minimums and maximums are reflected in actual and estimated standard deviation. One can see that the actual standard deviation is three to four times as high as the estimated one. If one compares the actual and estimated minimums and maximums, it is hard to decide which year has the best results. According to standard deviation comparison however, the best fitting appears to be for 2007 and 2011 (when maximums in actual data were the lowest), which are also the years with the highest R^2 values as presented in Table 11.

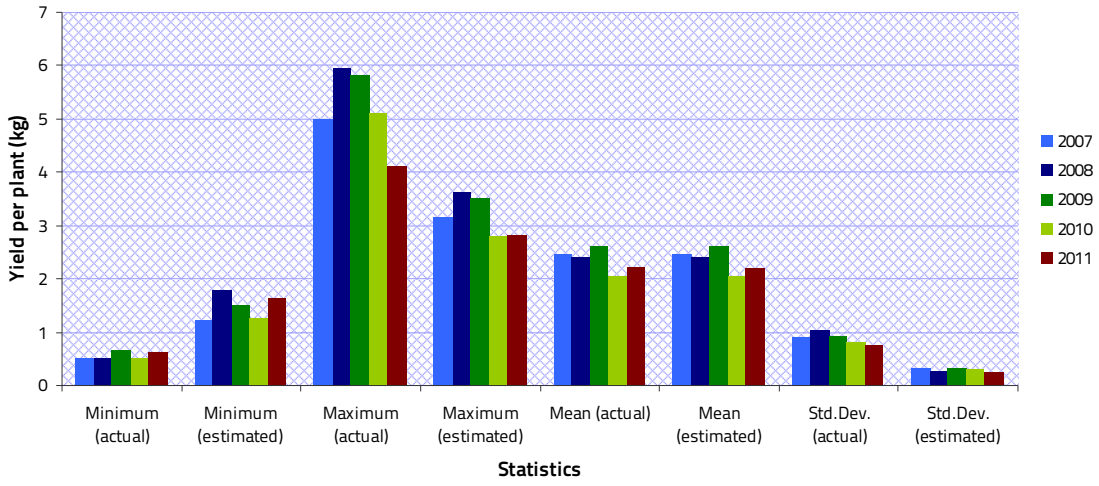


Figure 27. Statistical comparison of actual and GWR estimated yield (2007 – 2011, Rebula).

Figure 28 depicts the maps of local R^2 for Rebula variety for 2007 – 2011. Again, one must be careful when reading the maps because the ranges of legends differ. One can see that there appears to be a pattern in all maps as some areas have higher local R^2 values than others. Moreover, the distribution of low and high local R^2 values is similar in some maps. For example, the maps for 2007 and 2009 are very similar, as they both have high local R^2 values in the northern part of the area and low local R^2 values in the southern part of the area. Maps for 2008 and 2010 are similar as well as they both have high local R^2 values in the north-eastern part and the lowest R^2 values in the western part of the area. The map for 2011 is completely different than the others. The highest local R^2 values are in the south-western part, but high values occur in the north-eastern part as well. The lowest values in 2007 are in the north-western part of the area. However, all of the highest local R^2 values are still very low and even in the areas with the highest local R^2 value the GWR model can not explain enough variance to be deemed feasible.

The following additional visualizations for Rebula variety for every year from 2007 to 2011 are available in Appendix C:

1. plot of GWR estimation on actual yield per plant for Rebula variety,
2. plot of residuals to estimated values for Rebula variety,
3. local R^2 histogram,
4. residuals histogram for Rebula variety,
5. map of residuals for Rebula variety.

When one observes the visualizations in Appendix C, one can summarize GWR characteristics for estimation of Rebula variety yield per plant for years 2007 – 2011 thusly:

1. there is positive correlation between estimated and actual values of yield per plant for all years while the estimation for 2007 seems to fit the most the actual yield; there is a lot of variance in estimations though and the estimation of extreme values is unsatisfactory,
2. there appears to be no relation between residuals and estimated values,
3. there is no pattern visible when comparing the local R^2 histograms, though in 2008 and 2009, when the models accuracy is the lowest, almost all records are explained with less than 0.1 R^2,
4. the residuals histograms are very flat, meaning that the accuracy of the model is low; the range of residuals varies considerably, from approximately 3.2 (year 2011) to approximately 4.8 (year 2008) kg,

5. there appears to be no pattern on a map of residuals, though variations in range of residuals, already identified in histograms, are obvious.

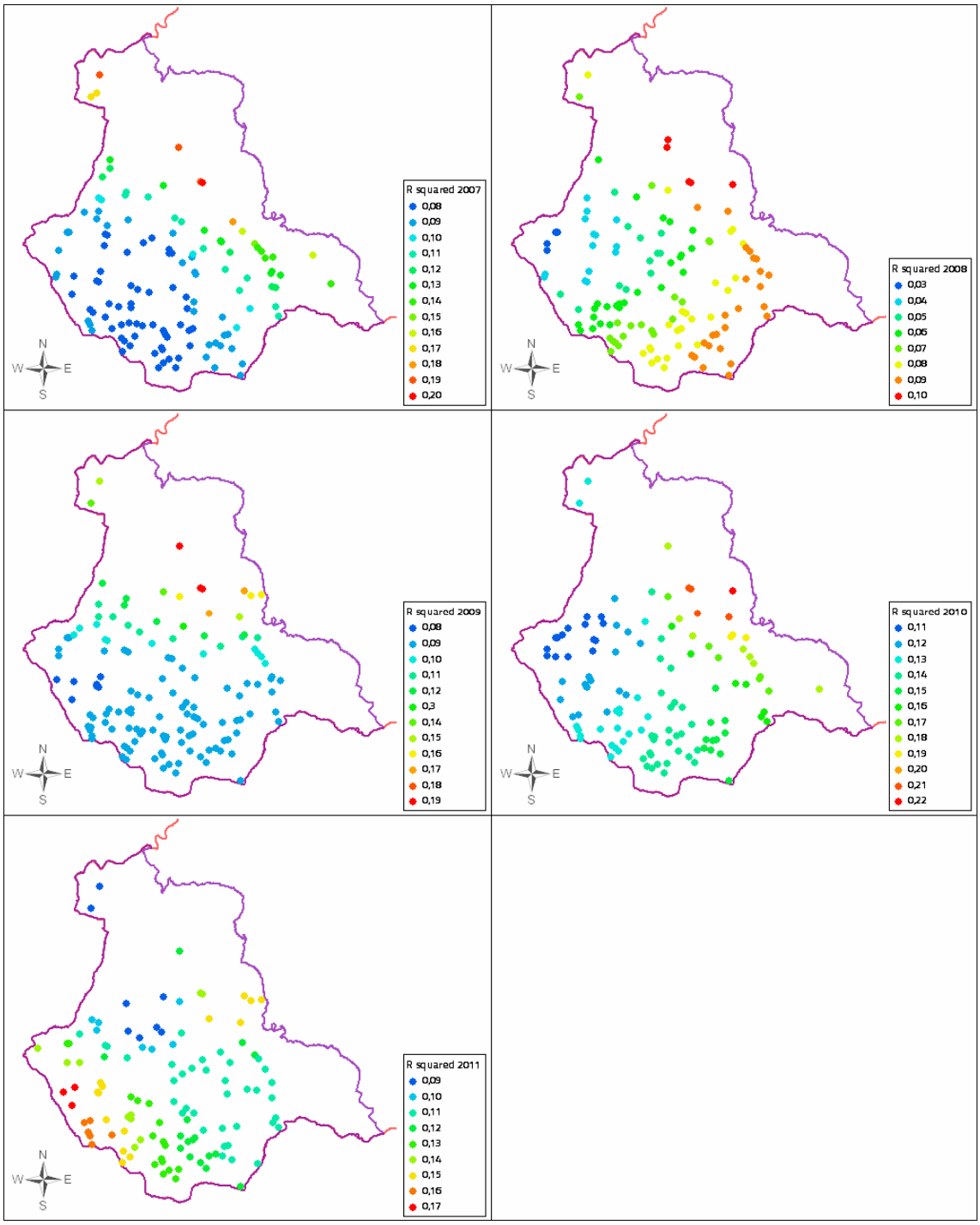


Figure 28. Local R^2 values of GWR (2007 – 2011, Rebula).

Discussion

The results of GWR proved that local regression method using explanatory variables as used in this research can not predict yield. The highest explained variance reached during the five years time period of this research does not even reach the threshold which would define it as weak. Some comparisons and conclusions can be made nevertheless.

The first comparison that should be discussed, is one regarding R^2 values obtained by the OLS and GWR methods. First of all, it must be mentioned that none of the methods are able to explain

enough variance to be considered feasible. However, GWR eventually explains a higher ratio of variance than OLS, particularly in the case of all data. The results for Rebula variety on the other hand are not as significantly in favour of GWR. This is because even though GWR explains more variance than OLS, OLS proves to be superior according to AICc scores. However, one can speculate that more Rebula data records could eventually result in a better assessed GWR model.

Furthermore, there appears to be a spatial pattern in the distribution of high and low local R^2 values that is in some cases repeated over the years. If R^2 values were significant, this would be an issue for further research, as it could provide more information about the explanatory power of variables. For example, one could find out if the prediction is more accurate in the lowlands than in a hilly terrain. One could also find out which variables ad up the most of explained variance of a particular data record. Similar findings could be also searched for in maps of residuals in case they would actually show some patterns. In our research however this was not the case.

The comparison between the prediction capabilities of GWR for entire yield and Rebula yield can be made as well. Figure 29 depicts the summary of GWR models for entire yield and for the yield of the Rebula variety for 2007 - 2011. As seen in the figure, entire yield is in general explained better than yield of Rebula variety. In 2008 and 2011 however, the variation of Rebula variety yield is explained better. Furthermore, in 2008 approximately the same ratio of variance is explained for entire yield and for Rebula variety. Finally, in years 2009 and 2010, the difference in favour of entire yield is considerable, as It reaches approximately 0.1 of R^2. One can again speculate that more numerous and denser Rebula data records could result in a more accurate GWR model. Nevertheless, based on the relations of R^2 for entire yield and for the yield of the Rebula variety, one can assume that the consideration of grape variety type in modelling would not be likely to contribute to an increase of R^2 of the GWR model.

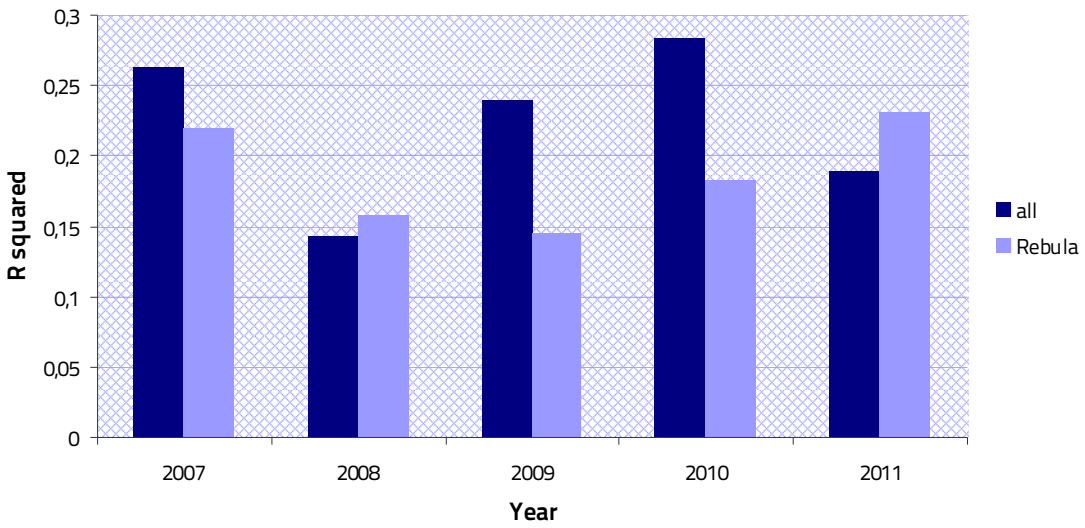


Figure 29. Comparison of R^2 of GWR yield estimations for all yield and for Rebula (2007 – 2011).

The search for possible causes of such poor results should first focus on data used in the research. Apparently, there are no global or local relations between explanatory variables and the dependent variable. By considering additional variables, this could have changed. Furthermore, optimisation of GWR settings might produce better results. However, the GWR settings for this research were selected after consulting literature and testing. Moreover, the most delicate setting, that is the bandwidth, was optimised by a software algorithm, taking into account a standard model assessment criterion (AICc).

5.4 SOM

Within this subchapter, the results of training of the 2009 dataset are presented first. The projection of the derived clusters onto geographic space follows. Next, the results of mapping the data from 2007 – 2011 onto the trained network based on 2009 data are presented. Finally, the relation between the clusters and the dependent variable is in focus. These issues are first presented for all data and followed by the presentation for Rebula variety data.

All data

The process of testing of SOMs for all data is depicted in Figure 30. One can see the range of tested parameters, as well as the ones eventually selected (dimension of 10x16; initial learning rate of 0.05; 60 000 iterations).

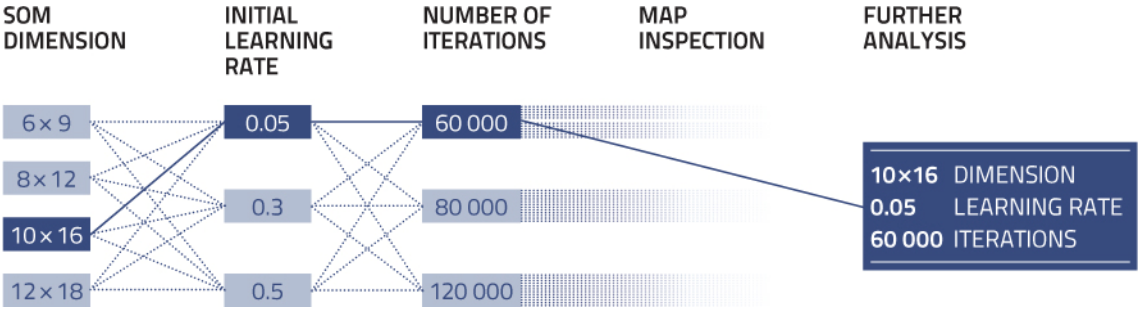


Figure 30. The process of discovering optimal SOM properties.

On the U-matrix illustrated in Figure 31 (left), the selected SOM is depicted. One can identify four clusters (neurons with low values delimited by neurons with high values), whose borders are in some parts barely visible. The cluster borders were identified visually in this research despite the fact that there is an option for automatic clustering available in R kohonen software package. Automatic clustering however proved to be unsuitable for our cause, as it often assigned the same cluster even to those neurons, which were not direct neighbours. However, the visually identified clusters that were used in further analysis are shown in Figure 31 (right).

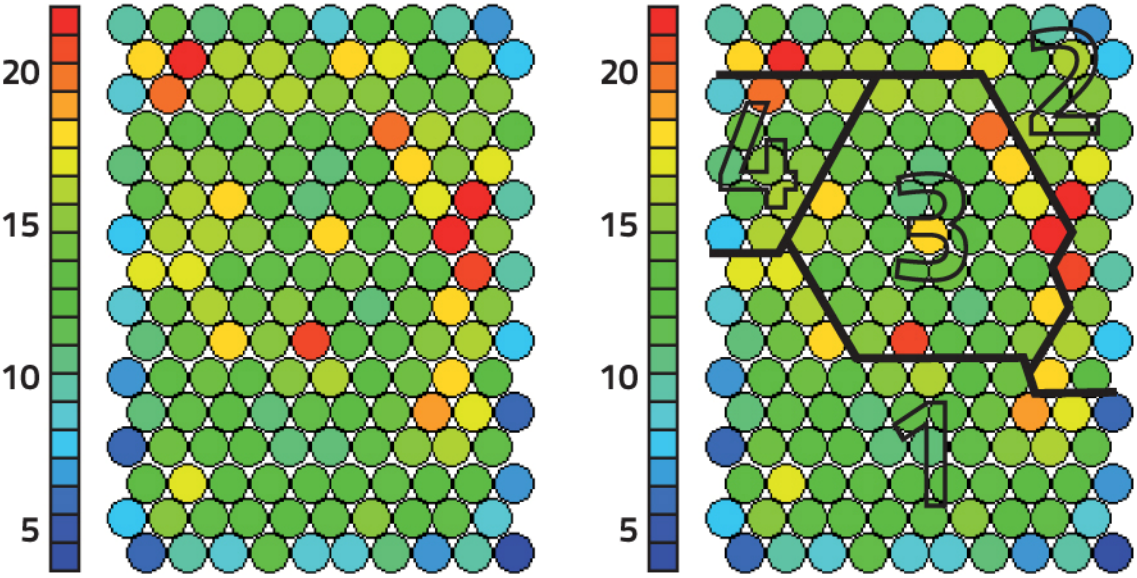


Figure 31. U-matrix of the derived SOM (left) and clustered U-matrix of the derived SOM (right).

The component planes belonging to the U-matrix shown in Figure 31 can be observed in Figure 32. On component planes one can identify areas where variables have high or low values. If borders between low/high values in the component plane of a certain variable coincide with a certain cluster border on U-matrix, then that particular variable is (at least partly) responsible for delimitation between the clusters in question. By comparing the component planes with the U-matrix in our case, one can see that Y_PLANTING has influenced the clustering the most, as it has strongly influenced the delimitation between the second cluster and the other clusters. Other variables appear to have considerably less influence on clustering, though some similarity between cluster assignation and the distribution of high/low values on component planes can be seen. For example, the similarity between SLOPE_AVG and EXP_AVG component planes is obvious in lower right part of the plots. Apparently, the vineyards on flat land (at the bottom of valleys) with zero slope and consequently zero exposition were mostly mapped to that part of SOM.

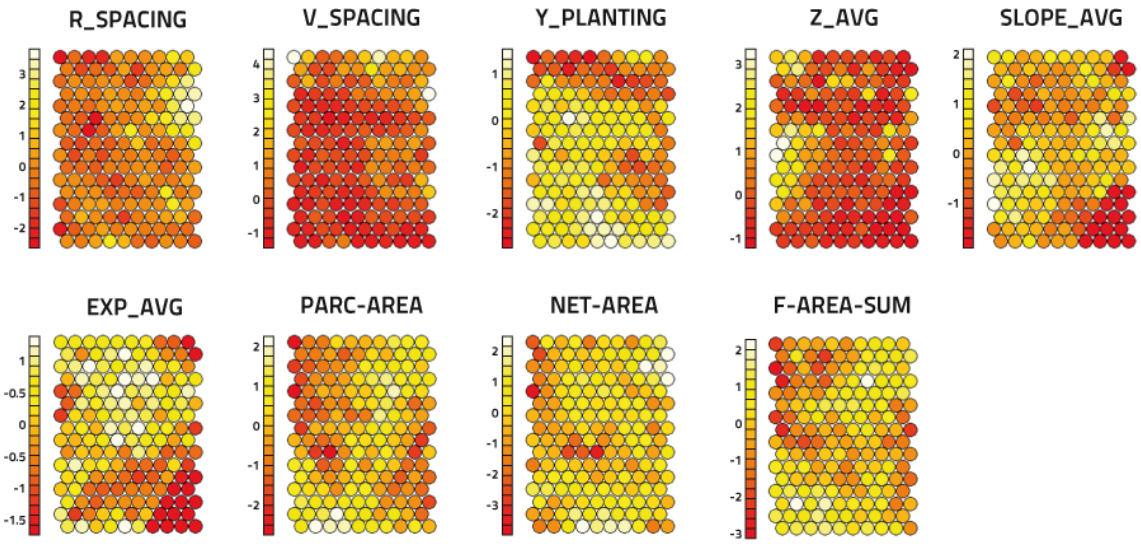


Figure 32. Component planes of the derived SOM.

In Appendix D one can observe additional plots describing this SOM, namely the count of occurrences plot, mapping quality plot and distance change plot. The count of occurrences plot shows the number of data records assigned to a particular neuron. If one compares the plot to the U-matrix, one can see that the count of data records assigned to neurons close to cluster borders is usually low, while the count of data records assigned to neurons in central parts of the clusters is usually medium or high. This is good, because apparently most of the data records assigned to a particular cluster are similar to other data records within that same cluster. In the next plot, the mapping quality plot, low values indicate better mapping quality. In our case most neurons have good or medium mapping quality, while there are few neurons with relatively bad mapping quality. Mapping quality of SOM that is presented here was somewhat worse than mapping quality of SOMs with the same properties, but higher starting alpha values. That means that though the clusters were most visible on this particular SOM, the similarity of data records within neurons was better in some of the other tested SOMs. The final plot in Appendix D depicts the change of mean distance to the closest unit during the training. Apparently the number of iterations was sufficient, as the mean distance was barely changing in the last part of the training process.

Table 12 shows the number and ratio of data records that were clustered in four clusters. The results between 2007 and 2011 and within each of these years can be observed. One can see that the data records are quite unevenly distributed among the clusters. The first cluster includes approximately 45 % of the data records, while the third cluster includes approximately 30 % of all data records. The remaining data records, less than one quarter of the whole dataset, are in the second and the fourth cluster. The latter one is with less than 8 % data records the smallest out of

all four clusters. The ratios of data records in each cluster are approximately the same during the whole period. The projection of clustering to a temporal dimension therefore indicates that the clustering is stable during the whole research period. There appears to be no big change in the data during the time period in question.

Table 12. Number and ratio of data records per cluster (2007 – 2011).

Cluster	2007	2007 %	2008	2008 %	2009	2009 %	2010	2010 %	2011	2011 %
1	535	45.8	529	45.8	549	44.8	529	45.9	507	46.8
2	205	17.5	197	17.1	219	17.9	195	16.9	175	16.1
3	351	30	340	29.5	363	29.6	339	29.4	318	29.3
4	78	6.7	88	7.6	94	7.7	89	7.7	84	7.8

The projection of clustered data records (2009 data example) into geographical space is shown in Figure 33. One can see that all clusters are more or less evenly spread over the whole area covered with vineyards. None of the clusters appears to be absolutely dominant in any part of the area. First cluster though is strongly represented in the southern part, but on the other hand, almost half of the data records were assigned to that particular cluster (Table 12). One can also notice that the groups of vineyards, which are connected to patterns in the southern part of the area, usually contain vineyards that belong to the first cluster. The patterns appear to follow the valleys of flat relief, which is in fact present in that part of the area. As noticed on component planes in Figure 32, most of the data records with low slope and exposition have indeed been assigned to the first cluster. The projection of clusters to a geographical space therefore proves that there is some correlation between the clustering result and the geographic location of clustered data records, but the patterns are local and there are no larger areas assigned to one particular cluster.

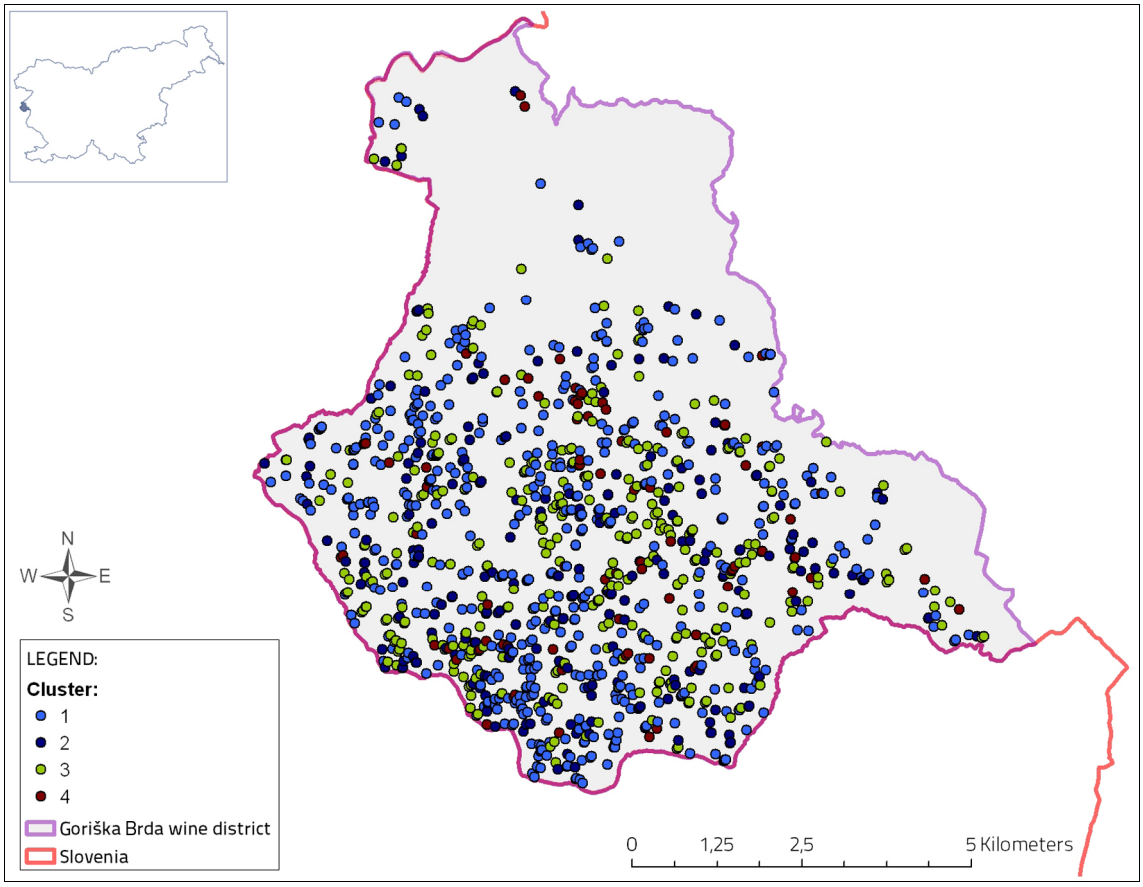


Figure 33. Geographical distribution of SOM clusters (2009 data).

The relations between the clusters and the dependent variable are shown in Figure 34 and Figure 35. As seen from Figure 34, there appears to be a certain correlation between the dependent variable and clusters during 2007 – 2011 period. For example, mean yield of data records assigned to the first cluster is always approximately 0.5 kg lower than mean yield of data records assigned to the second cluster, where yield is highest. With exception of 2007, yield is always highest in the second cluster, which is followed by the third, fourth and finally the first cluster. One can relate the content of Figure 34 to the component planes of SOM depicted in Figure 32, where the year of planting was identified as the most important variable in forming of the second cluster. According to the component plane of that variable we can therefore estimate that in general older vines appear to produce more yield. Another important remark about the figure below is that, though there apparently is a connection between the cluster assignation and yield, the differences between mean yields between clusters are quite small. In order to appropriately asses the relation between mean yield per plant and cluster assignation, one therefore has to inspect additional statistics.

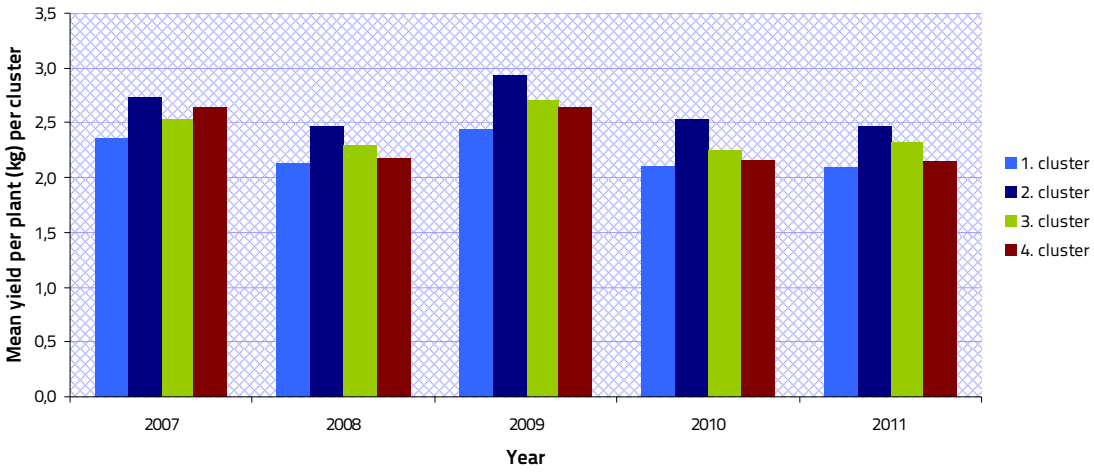


Figure 34. Mean yield per plant per cluster (2007 – 2011 data).

In Figure 35, the range of yield of data records within clusters is depicted. One can see that the range of yield per plant within the cluster is in most cases around five kg per plant, the exception however is the range of the fourth cluster in 2008 and in 2011, with a range of slightly above 3 kg. If one accounts for another fact, that is that the range of actual yield per plant in our research is always between 0.5 kg and 6 kg, the differences between the clusters that are seen in Figure 34 appear even less meaningful. In fact, they are too small to clam that we could estimate yield with clustering.

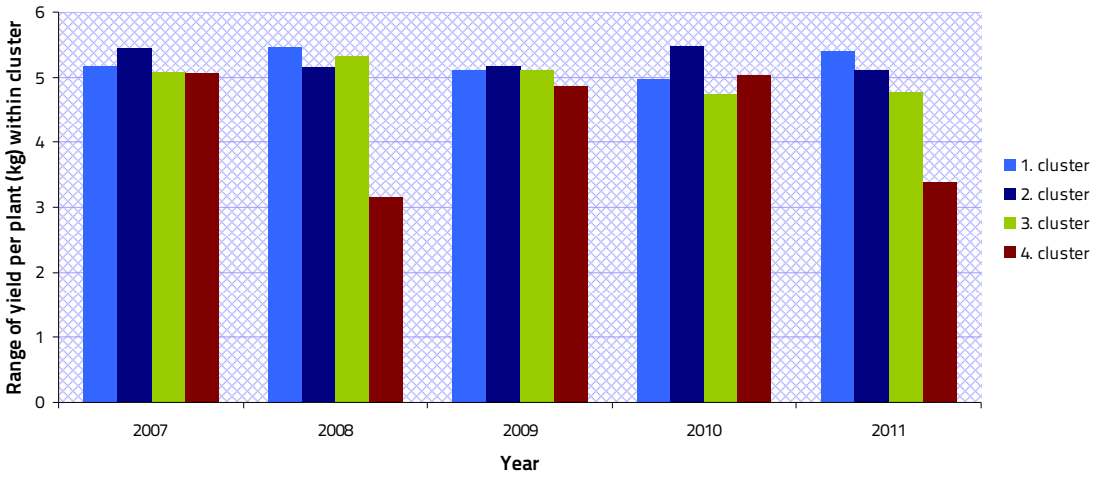


Figure 35. Range of yield per plant within clusters (2007 – 2011 data).

Another statistic that indicates that we can not estimate yield with clustering (in this research) is standard deviation of variable within clusters, which is shown in Table *13*. One can see that standard deviation within the cluster is in all years for all clusters between 0.74 kg and 1.18 kg, which is more than is the largest difference of yield per plant between the clusters. On can also see that it is the lowest in the fourth cluster which has the fewer number of data records, while it is surprisingly the highest in the second cluster, which has the second fewest the number of data records.

Table 13. Standard deviation of yield per plant (in kg) within clusters (2007 – 2011).

Cluster	2007	2008	2009	2010	2011
1	0.94	0.94	0.93	0.87	0.91
2	1.15	1.12	1.18	1.17	1.02
3	0.93	0.95	0.94	0.92	0.85
4	0.87	0.78	0.83	0.83	0.74

Rebula variety data

The testing process and the selected settings of SOM for Rebula data are depicted in Figure 36. As one can see, the largest of the two tested dimensions (6x9), the medium option of 0.3 initial learning rate and the fewer tested iterations (20 000) resulted in the most clearly clustered SOM.

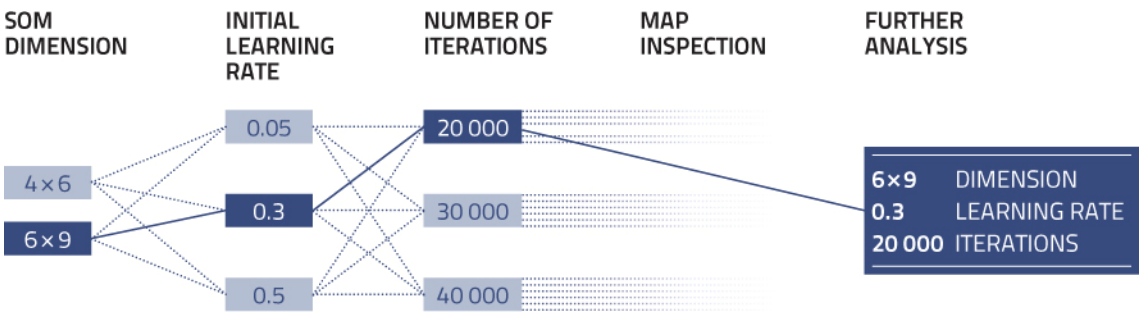


Figure 36. The process of discovering optimal SOM properties (for Rebula data).

The derived U-matrix and its clustering can be observed in Figure 37. Three clusters can be visually identified from the plot. In this case, the borders between the clusters are more obvious than in the case of all data. One can also see that in the case of Rebula, the size of the clusters (in terms of the number of neurons) is more similar.

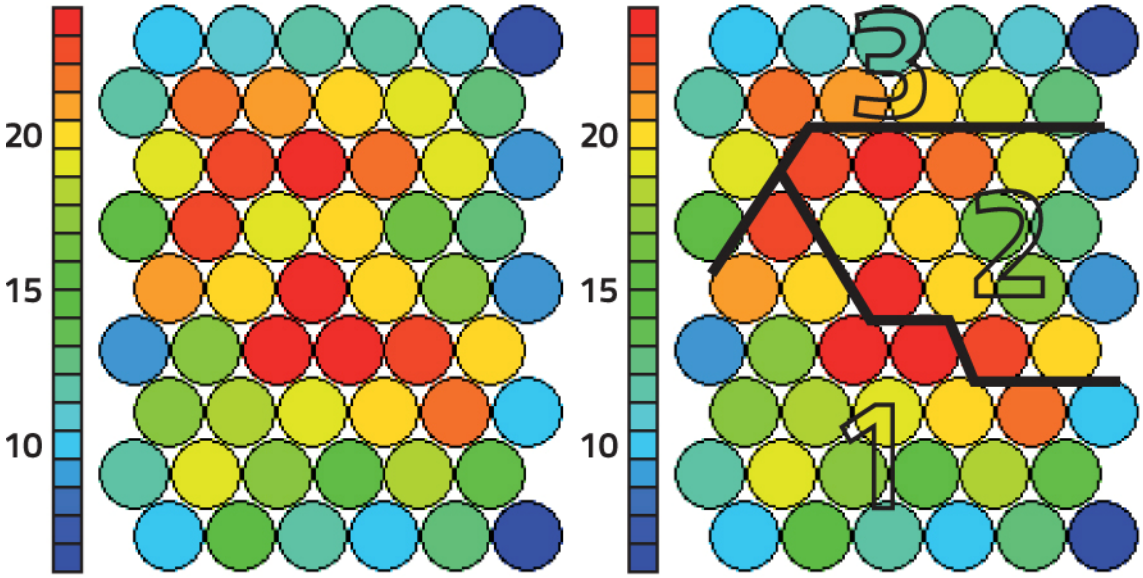


Figure 37. U-matrix of the derived SOM (left) and clustered U-matrix of the derived SOM (right) (Rebula data).

The component planes depicted in Figure 38 indicate that the areas with low or high values of individual variables are not coherent. By this we mean that the areas with low or high values in a particular component plain are not as uniform as for example in case of all data, depicted in Figure 32. Nevertheless, by observing component planes below and the U-matrix above, one can claim that none of the variables plays a particularly important role in clustering as cluster borders do not follow borders in any of the component planes.

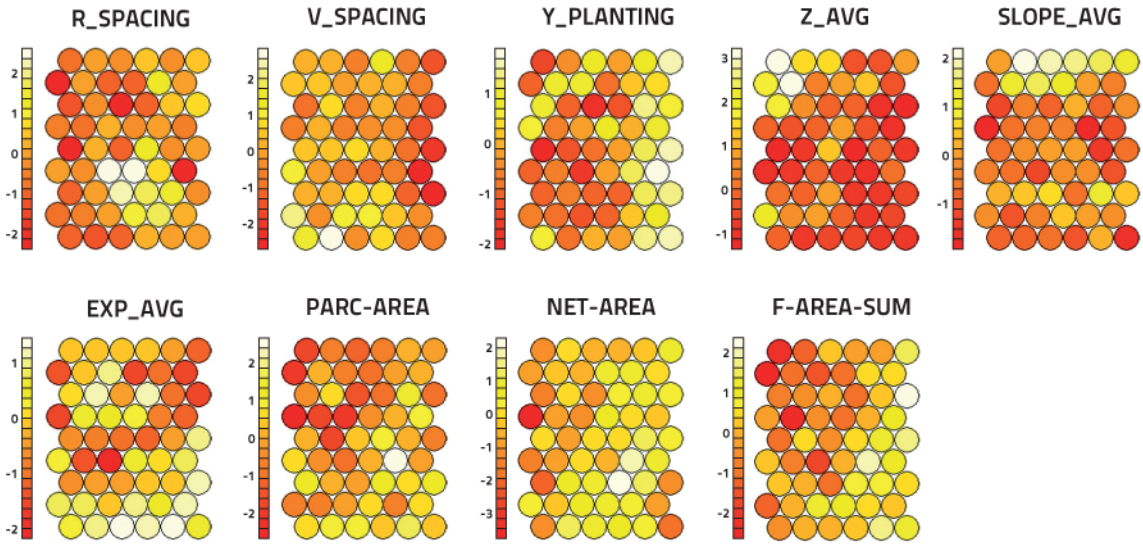


Figure 38. Component planes of the derived SOM (Rebula data).

In Appendix E, additional plots that describe this SOM can be observed. The count of occurrence plot is interesting because it shows that the SOM might even be regarded as overfitted, because more than half of the neurons describe only one data record. On the other hand, one can observe that the neurons, to which the most data records are assigned, lie within the central part of the clusters. Moreover, borders between clusters almost in all cases follow the neurons where only

one data record is described by a neuron. These facts indicate that the clusters should be homogenous. The mapping quality plot was expected to show a high number of neurons of very high mapping quality, as many neurons contain only one data record. The mapping quality of neurons varies and is in most cases worse with increasing number of data records assigned to a neuron, which is expected. What is not expected however is that not all neurons that contain single data record have the best possible quality. Finally, the plot of training progress, showing mean distance to the closest unit during the training, shows that the number of iterations for this SOM was sufficient because the mean distance hardly changed during the last part of the training process.

The change in number and ratio of Rebula data records, belonging to a particular cluster between 2007 and 2011, can be observed in Table 14. One can see that the first cluster has approximately the same ratio of data records throughout the period of research. However, the share of data records in the second cluster is decreasing, while the share of data records in the third cluster is increasing. By taking into account the fact that the third cluster represents many of the data records with the highest altitude and slope (Figure 38), one might for example argue that the ratio of vineyards planted with Rebula is increasing in areas of higher and steeper relief.

Table 14. Number and share of data records per cluster (2007 – 2011, Rebula).

Cluster	2007	2007 %	2008	2008 %	2009	2009 %	2010	2010 %	2011	2011 %
1	58	45.0	57	44.5	63	46.3	54	44.3	52	44.4
2	47	36.4	44	34.4	38	27.9	40	32.8	36	30.8
3	24	18.6	27	21.1	35	25.7	28	23.0	29	24.8

Rebula data records as clustered using SOM are projected to geographical space in Figure 39. One can see that the data records belonging to the second cluster are evenly distributed over the area. On the other hand, the data records of the first cluster are more common in the southern part of the area, while the data records from the third cluster are more common in the northern part of the area. By observing component planes in Figure 38, one can see many neurons that represent the highest altitude (Z_AVG component plane) and neurons that present the highest slope (SLOPE_AVG component plane) are assigned to the third cluster. More frequent occurrence of this cluster in central, orographically higher and hillier part of the area is possibly caused by this.

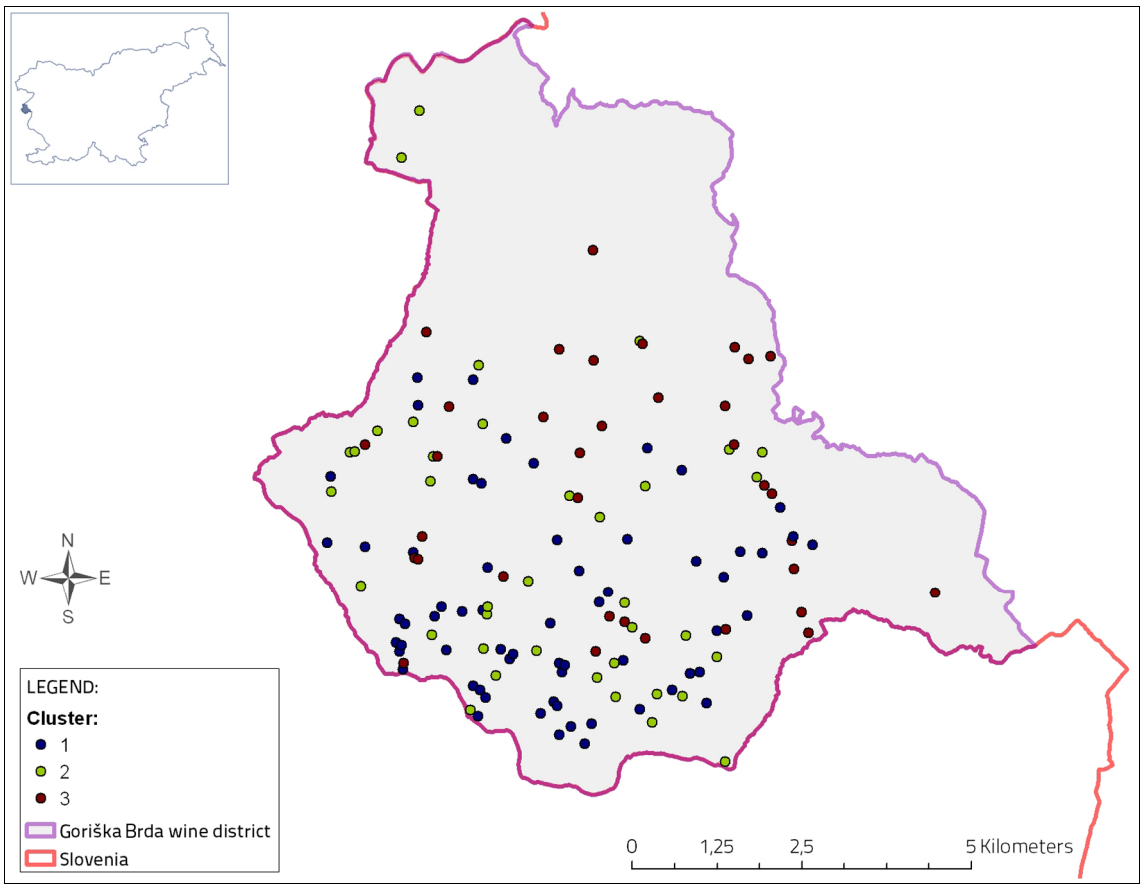


Figure 39. Geographical distribution of SOM clusters (2009 Rebula data).

The relation between clusters and yield per plant of Rebula variety from 2007 – 2011 can be observed form Figure 40 and Figure 41. Relation of mean yield per plant to clusters is shown in Figure 40. The mean yield of the data records assigned to the first cluster is the lowest during the whole period. The highest mean yield in 2008, 2009 and in 2011 is in the second cluster, while the highest mean yield in 2007 and 2010 is in the third cluster. One can see that the difference in mean yield of clusters varies between years. In 2008 for example, the difference between mean yield of the first and the second cluster is almost 0.5 kg, while in 2010 the mean yield of all clusters is almost the same.

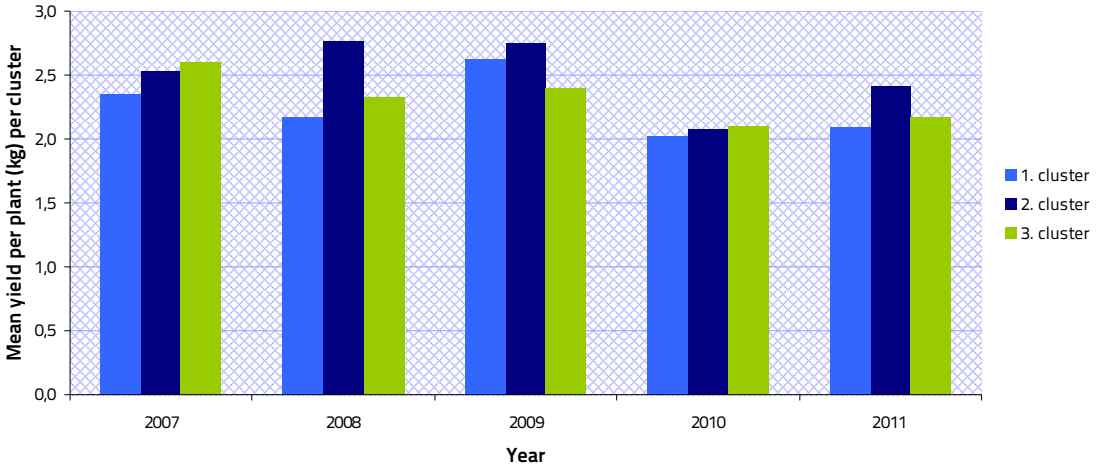


Figure 40. Mean yield per plant per cluster (2007 – 2011 Rebula data).

Figure 41 shows the range of Rebula variety yield per plant within clusters from 2007 - 2011. One can see that with the exception of 2011 there are big differences between ranges of yield per plant between clusters of the same year. One can also see that there is no rule in the range size of particular cluster from year to year. Such variation however can be the consequence of the small number of data records used. Nevertheless, the range of yield per plant of all clusters and all years is large, which makes this approach not suitable for accurate estimation of yield.

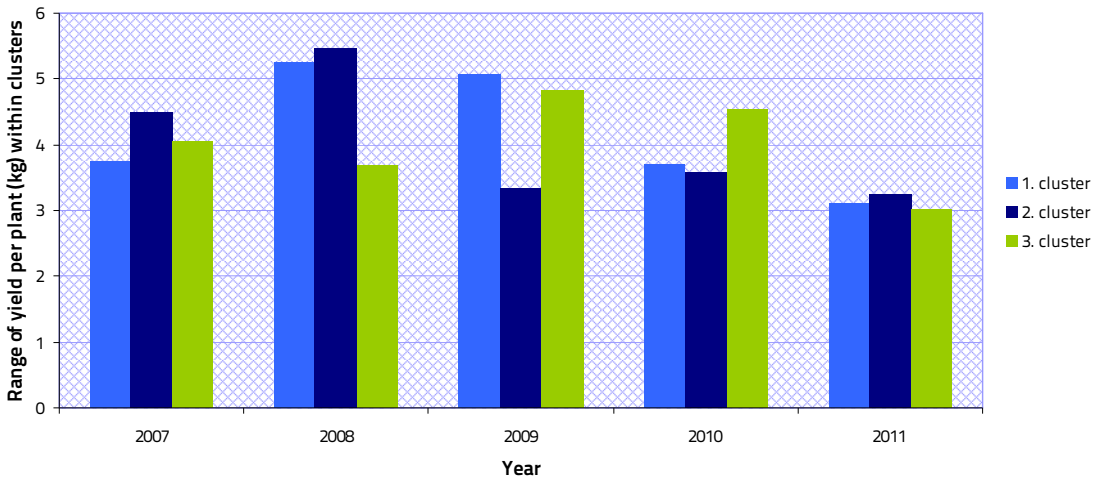


Figure 41. Range of yield per plant within clusters (2007 – 2011 Rebula data).

Standard deviation of yield per plant within clusters is depicted in Table 15. One can see that the standard deviation within the cluster is in all cases higher than the differences in mean yield between the clusters (depicted in Figure 40). All these statistics prove that we can not estimate Rebula yield with clustering (in this research).

Table 15. Standard deviation of yield per plant (in kg) within clusters (2007 – 2011, Rebula).

Cluster	2007	2008	2009	2010	2011
1	0,88	0,99	0,98	0,72	0,74
2	0,91	1,16	0,90	0,90	0,88
3	0,93	0,87	0,80	0,90	0,68

Discussion

The inspection of testing results did not indicate strong preference of SOM properties. No clear clusters could be identified from the obtained U-matrices, regardless of the combination of dimension, starting alpha value and number of iterations. This was the case particularly for all data and to quite some extent for Rebula data as well. The testing indicated that where the clusters occurred, they could have occurred randomly rather than because they would actually exist as a combination of explanatory variables. However, phisiogeographical variables, particularly altitude, appeared to be an important variable in clustering for both the whole of the data and Rebula data.

Poor results in further analysis were therefore expected. The temporal component appears to have no particular influence on any of the results that are of interest for this research. These findings apply to both the experiments conducted with all of the data and with data of the Rebula variety. Projection of clusters into geographic space indicates that there is no strong geographic

pattern in the location of clusters. Furthermore, according to the results there is no particular relation between the clusters derived from the SOM and the dependent variable, being yield per plant in kg, especially when considering range and standard deviation of data records assigned to clusters.

There are a few explanations for the poor performance of SOMs in our research. First, it is possible that SOMs are not suitable for our research problem. Second, it is possible that SOMs were not applied correctly. Third, it is possible that there actually is no pattern in the data. The following might be because the most important variables to explain the variation are missing.

The first possibility, that the method is not appropriate for our research problem, is unlikely. SOMs were so far applied to solve a number of different clustering problems and were in many cases successful. Therefore, in our opinion there should be some results when applying SOM in our research in case there actually is a clustering pattern in the data and of course in case the method is applied correctly and the data is prepared appropriately. Other data exploration methods might yield better results, but the comparison of clustering methods was not the goal of this research.

The second possibility, that the method was not applied correctly, is always a possibility in scientific work. However, the literature research of SOM training and the review of SOM results were done thoroughly. A number of testing experiments, based on settings as recommended in literature were performed in order to find the best possible SOM parameters for our case. The method was applied according to instructions in the manuals and literature. Furthermore, alternative SOM software, GeoSOM was tested as well, but eventually abandoned for practical reasons, connected with results display and reviewing capabilities. Consequently, because of thorough research and testing, we believe that the possibility of incorrect application of the method was minimized.

Finally it is possible that there really is no pattern in the data that were used in the research. This is in our opinion the most realistic reason for not achieving the desired results. For example when iterating the experiments using the same SOM settings, but different starting seed (or even by ordering a data by different column when importing it) the resulting plots were very different. According to literature however this is somehow expected. U-matrices deriving from the same data and settings, trained using different seed, can have different shapes of clusters, but should on the other hand have the same data records assigned to the same number of clusters. In our case the differences between such U-matrices were immense. For example in case of applying two iterations with the same settings but with different starting seed, in the first U-matrix some clusters were visible, while in the second one, no clusters were visible at all. In case the clusters appeared, they seem to have appeared randomly.

It must however be mentioned that initially more variables were meant to be applied to SOM data mining than were eventually applied. Meteorological variables, soil properties, training system type, rootstock type and grape variety type were acquired as well, because they are important from the perspective of this research, as they have an influence on vineyard yield. The inability to use meteorological data is described in section 5.1. The inability of use of the remaining variables is their qualitative data type and their characteristics. According to literature, SOMs were successfully applied using qualitative (Ritter & Kohonen, 1989) or even qualitative and quantitative data (Vesanto, 1997) in the same research. In our case, the usage of 1-of n coding was considered for preparation of qualitative data (Vesanto et al., 2000). It was eventually not applied in the research because of the characteristics of the data used in the research. For example, there are 18 different training system types and 28 different grape varieties within the data. Applying the coding to all possible types would eventually result in domination of these variables in SOM.

Results of projection of clusters to geographical space and of explanation of a dependent variable using SOM clustering should be discussed as well. Bad clustering results already initially considerably decreased the possibility of deriving satisfying results of these experiments.

With regards to the projection of clusters to geographical space it should be stressed that certain limitations of this approach were known beforehand. For example, the variables such as average slope, average exposition and average altitude vary immensely in a hilly type of relief, which is the case in our study area. One could therefore expect that these explanatory variables will contribute to geographic non-homogeneity of clusters. Explanatory variables such as distance between rows and distance between vines, which are in many cases dependent on the terrain properties, could add to geographic non-homogeneity of clusters as well. On the other hand, variables such as year of planting, the sum of the farm's vines, area of vineyard covered with vine in question, etc, could be more spatially coherent and could contribute to the occurrence of spatial patterns from clusters in case they existed. Moreover, some other variables, as mentioned above, that were initially planned to be applied in the research, could have contributed to formation of spatial pattern of clusters as well.

Finally, the experiments showed that the dependent variable, yield per plant, was not clustered according to its quantity. According to the results, the distribution of for example data records with high yield per plant was apparently assigned to clusters randomly. The reasons can be found in SOM clustering problems described above and in some other issued of predicting yield per plant that are described in general discussion.

5.5 Comparison of OLS and GWR results with meteorological characteristics

The comparison of 2007 – 2011 OLS and GWR R^2 results with meteorological characteristics is presented in Pearson's correlation matrix in Table 16. For the purpose of presentation, the fields that we are focused on have bolded numbers, while the correlations interesting from our perspective are highlighted green or yellow, depending on the strength of correlation.

The correlation matrix shows that mean temperatures are in most cases positively correlated with the accuracy of the model. The correlation is the highest for OLS model for Rebula, though it is still relatively weak. Consequently it appears that in general if mean temperatures between April and including September are above average, the models appear to be slightly more accurate.

Table 16. Comparison of OLS and GWR results with meteorological characteristics (2007 – 2011). Source: SEA

	Mean temperat.	Precipitation	Sun hours	OLS R^2	GWR R^2	OLS R^2, Rebula	GWR R^2, Rebula
Mean temperature	1	-0.917	0.779	0.292	-0.076	-0.368	0.23
Precipitation	-0.917	1	-0.739	0.1	0.314	0.64	-0.217
Sun hours	0.779	-0.739	1	0.315	-0.016	-0.142	0.787
OLS R^2	0.292	0.1	0.315	1	0.492	0.618	0.276
GWR R^2	-0.076	0.314	-0.016	0.492	1	0.875	0.138
OLS R^2, Rebula	-0.368	0.64	-0.142	0.618	0.875	1	0.225
OLS R^2, Rebula	0.23	-0.217	0.787	0.276	0.138	0.225	1

One can also see that there is some interesting correlation between precipitation amount and modelling accuracy. Precipitation quantity is positively correlated to the accuracy of GWR model for all plants and furthermore, it has strong correlations with the accuracy of the OLS model for Rebula. Apparently, the increase in precipitation quantity between April and including September increases the prediction power of most models as well.

Finally, the number of sun hours appears to be loosely correlated to accuracy of the OLS model for all varieties and in particular with the accuracy of the GWR model for Rebula variety. The latest model's accuracy is strongly correlated with the number of sun hours at the coefficient of 0.787.

However, taking into account weak or random correlations of OLS and GWR models for the whole yield and Rebula variety, one should not pay much attention to the correlations identified above. That is because it is possible that a slight difference in strength of models might result in different correlation of model with meteorological characteristics.

By inspecting the remaining values in Table 16 one comes across another interesting relationship, which is not the focus of research here, but should be mentioned because of its very high correlation of 0,875. The better the GWR model for entire yield, the better the OLS model is for the yield of the Rebula variety.

5.6 General discussion

As highlighted in previous sections, none of the yield estimation methods provided suitable results. Possible causes for this were discussed after presenting the results. Here, by summing up the main results and discussions, we can point out the difficulties that may have contributed to the poor results of this research. These are the following:

- the absence of a clear yield pattern in the dataset,
- the suitability of the explanatory variables,
- the data preparation and parameterization of methods.

The absence of a clear yield pattern in the dataset

In order to find a pattern in the data, this must exist. On the one hand one could claim that when dealing with yield as dependent variable, there are always factors that affect its quantity; one just has to identify them. On the other hand, one can question the applicability and costs of the method that would account for all these factors. In our research, for example, the intention was to use only existing data. Such data is, in most cases, produced or collected for purposes other than yield estimation.

Furthermore, data pre-processing results indicated that the accuracy of the dependent variable is not optimal. For instance, 3-4 % of the data records had to be excluded from the analysis due to obvious mistakes in the farmer's declarations. This increases doubt in the accuracy of the remaining data, which was eventually used in the research. The dependent variable's data acquisition technique (farmers' yield declarations) confirms that mistakes are likely to occur. There are certainly many farmers that are capable to estimate yield (from produced vine) accurately. However, there are also some farmers that are not capable to do so or that for whatever reason declare false yield on purpose. By taking into account that all data records are not accurate, a decrease in the model's accuracy is expected. The verification of the acquired (dependent variable)

data on the field would increase the cost of this method's application considerably. To conclude, one can question the presence of an identifiable pattern in the data that was used in our research.

The suitability of the explanatory variables

The question of the appropriateness of the selected explanatory variables is relevant after reviewing their correlation with yield. Generally speaking, none of the selected explanatory variables are correlated to the dependent variable at a global level. We however do not perceive the absence of global correlations as critical, since some methods other than global regression were also used in the research. Possibly existing correlations could therefore have been found. Consequently, we do not think that the choice of the explanatory variables was inappropriate. According to the literature all of them might contribute to yield quantity.

Nevertheless, it is possible that the explanatory variables, which could explain the highest variation of the dependent variable, were a) not accurate enough, b) not acquired, or c) acquired and eventually not applied in the research.

The accuracy of some of the explanatory variables used in the research is questionable. One can first question all the variables that have been derived from the farmer's applications. Further, those variables calculated from DTM (slope, exposition and altitude) are accurate as average values of the vineyard, but do not account for the variation of the phenomena within the vineyard. Nevertheless, to our knowledge, it does not exist more accurate data than the one we used.

Regarding the second group, the variables that were not acquired, one can identify 'the farmer's goal" as the most crucial one. The information whether the farmers' goal is to produce the highest quality wine for sale, or medium / low quality wine for sale / his own consumption, could provide valuable insight. That is because certain farmers do target yield ratios due to influence of yield quantity on its quality. Such insight was intended to be included in the research by introducing socio-economic variables. Furthermore, additional meteorological data could prove useful as certain phenomena, such as hail, strong winds, etc can considerably decrease yield. On the other hand these phenomena usually have a local or regional influence and as such could be accounted for by applying local regression method, which we did.

Within the latter group, the variables acquired but eventually not used, one can expose meteorological data, vineyard properties and soil data. According to literature, meteorological data has the highest influence on yield out of these three. However, its influence on yield is more likely to differ from year to year than within the same year, especially on the area as small as the size used in this research. The vineyard properties (grape variety, training system and rootstock type) and soil type do have an influence on yield as well, but in relation with meteorological properties. Unfortunately, these variables can not be used because their of their data type (qualitative). In other words, there is a lack of spatio-temporal analysis methods able to simultaneously cope with qualitative and quantitative data. However, by considering the literature and results of this research, we assess that the application of these variables in the research, with exception of meteorological data, would most likely not have improved the results.

The data preparation and parameterization of methods

A critical researcher should always question whether the research methods were correctly applied. From this perspective, there are two critical parts a) data preparation and b) settings (parameters) used for the methods. In our case however, we believe that it is possible to deny such reason as a cause for poor results.

Data preparation usually includes various steps and some subjectivity in data preparation is always present in actions such as choice of transformation or outlier identification. Nevertheless, the data preparation in our research included actions as advised in literature and was therefore in our opinion appropriate.

Two of the methods applied were more complex from the parameterization perspective. The settings had to be set for GWR and SOM methods. They were set according to the guidelines found in literature and according to the testing results. Due to thorough literature research and large extent of the testing we can claim that the settings for these two methods as used for this research were suitable.

6. CONCLUSIONS

The main objective of this thesis was to use spatio-temporal data mining to estimate vineyard yield in Slovenia. More precisely, our goal was to build an after-harvest yield estimation model capable of estimating and verifying yield at any location. Such model would take natural conditions, vineyard properties, socio-economic variables and vineyards location into account. The research questions, formulated in section 1.3, define this goal more accurately. By answering the research sub-questions first, the main research questions will require less argumentation.

- *Q: Which available variables affect the quantity of yield most?*
- A: None of the available explanatory variables explains the variation of yield significantly.

This is the main result of the OLS and GWR regression methods and it is valid for the yield of all varieties and for the yield of the Rebula variety.

- *Q: Which of the two regression methods, OLS or GWR, estimates vineyard yield better?*
- A: GWR proved to be a better prediction method than OLS.

This proves that consideration of location increased yield estimation accuracy. The difference in prediction power however was small as both methods resulted in similarly poor absolute R^2 value. Again, this is the case for both, the entire yield and the Rebula variety yield.

- *Q: Can regression methods accurately estimate vineyard yield?*
- A: Neither OLS nor GWR can accurately estimate vineyard yield using the variables available for this research.

In fact, none of the regression methods can even approximately predict yield. The best prediction power of GWR for the entire yield is for 2010 (R^2 of 0.283), while the best prediction power for the Rebula variety yield is for 2011 (R^2 of 0.231).

- *Q: Can the SOM method successfully cluster available explanatory data?*
- A: The SOM method is not able to clearly cluster the available explanatory data.

No clear clusters are identified neither using all data nor using only Rebula variety data. Nevertheless, vague clusters were identified from trained SOMs for the purpose of further analysis.

- *Q: Is yield quantity reflected in clusters derived from SOM and do these clusters have a geographical pattern?*
- A: Yield quantity is not reflected in clusters nor do the clusters have a geographical pattern.

The variation of yield between the clusters was too small to be assessed as significant. Regarding the geographical pattern of the clusters, some of the clusters were more common in certain locations but no general patterns were distinguishable. This is the case for the entire yield as well as for the Rebula variety yield.

- *Q: What is the relationship between yield estimation accuracy and meteorological characteristics?*
- A: There is not a clear relationship between yield estimation accuracy and overall (yearly) meteorological characteristics.

Though there are certain correlations identifiable from comparison of OLS and GWR R^2 values and meteorological characteristics, they can not be identified as certain because the R^2 values of the regression are poor. This is the case for both, the entire yield and the Rebula variety yield.

- *Q: Can new findings be applied to research mid-season yield estimation?*

- A: New findings can partly be applied to research mid-season yield estimation.

No conclusions that would enable yield estimation are made therefore they can not be applied for mid-season yield estimation. On the other hand, one can claim that this research can help to avoid mid-season yield estimation research using the same methods and variables as here, because it is likely that it would not be succesful.

The answers to research sub-questions above outline the answers to main research questions, which are the following:

- *Q: Can spatio-temporal data mining as committed within this research provide the information suitable to construct a model for after-harvest yield estimation in Slovenia?*
- A: No, it can not.

The results show that the approach to spatio temporal-data mining as performed in this research did not provide the information suitable to construct a model for after-harvest yield estimation in Slovenia. The most likely reasons for that are the absence of a clear yield pattern in the dataset and the absence of appropriate explanatory variables.

- *Q: Can spatio-temporal data mining as committed within this research provide the most suitable method for constructing such a model?*
- A: No, it can not.

Because the variation of yield was not explained by any of the methods applied, none of them can be applied to construct a model for yield estimation.

- *Q: Can spatio-temporal data mining as committed within this research provide the basis for mid-season yield estimation in Slovenia?*
- A: No, it can not.

The inability to explain after-harvest yield variation proves that the basis for mid-season yield estimation can not be provided based on this research. On the other hand, this research does provide valuable information about the problems that are likely to be met in the case of mid-season yield estimation research.

One of the goals of this research, namely to build an after-harvest yield estimation model based on spatio-temporal data mining, was achieved. However, according to the results one can claim that spatio-temporal data mining, at least using the approach and the explanatory variables used in this research, can not accurately estimate or verify vineyard yield at a given location. Consequently, it was not possible to improve the current yield estimation methods in Slovenia. This was not a desired outcome of this research, but one has to conduct a research in order to confirm or deny a hypothesis.

Nevertheless, some of the results of this research are valuable to MAE. For example, it is now possible to identify those farmers, whose grape yield declaration or vineyards characteristics declaration is obviously incorrect. They can be identified by checking the dependent variable, yield per plant, outliers. Furthermore, the information about the quantity of yield per vine can be used to check whether the farmers, who produce high quality wine for sale, really do produce it considering yield quantity limitations required by legislation. From this perspective one can claim that the research provided useful tools for improvement of the accuracy of MAE registries. The registers accuracy could be improved by requireing corrected declarations from those farmers, whose declarations would be identified potentially incorrect.

The conclusion that something can not be done in a specific way, as it was the case in this research, can be valuable information for further research. It can even be a challenge for others to search for a different approach to solve the same problem. A few recommendations can be given

from the experiences gained when through this research. First, if the same approach is to be taken, additional explanatory variables should be searched for (farmers goal, farmers declared wine quality, relative altitude, ...) and as much errors as possible should be removed from the existing explanatory variables. Second, if yield is to be estimated at the level of vineyards for (mainly) small scale estimation and verification purposes, it would be feasible to perform such research on a smaller scale as a pilot project first. It could be applied in a farm (or few farms) level, to control the accuracy of the input data and to consequently more accurately identify and quantify the relations between the dependent variable and explanatory variables. Third, if yield is to be estimated at the level of the country or a region for (mainly) monitoring or policy making purposes, we think that it would be feasible to choose a larger spatial unit than vineyard (to add up the results), for example a wine district or wine sub-district. However when using such an approach one should reconsider the explanatory variables that were used in our research (some can be averaged, while some could not be used), and additional variables, such as meteorological data, that could possibly be applied. The approach based on a larger spatial unit and a larger research area could reveal certain variability in spatial distribution of yield, which may not have been identified when applying our approach.

REFERENCES

- Abbazia. (2012). Abbazzia di Rozazo web page. Via http://www.abbaziadirosazzo.it/home_abbaziadirosazzo.php?n=45&l=sl (accessed 05-05-2012).
- Andrienko, G. et al. (2006). Mining spatio-temporal data. Journal of Intelligent Information Systems, 27(3), pp. 187–190.
- Baluja, J. et al. (2012). Assessment of the spatial variability of anthocyanins in grapes using a fluorescence sensor: relationships with vine vigour and yield. Precision Agriculture, 13(4), pp. 457–472.
- Belec, B. et al. (1998). Slovenija: pokrajine in ljudje. Ljubljana: Mladinska knjiga.
- Blom, P.E. & Tarara, J.M. (2009). Trellis Tension Monitoring Improves Yield Estimation in Vineyards. Hortscience, 44(3), pp. 678–685.
- Bole, D. (2010) in Perko, D. & Zorn, M. (2010). Geografski informacijski sistemi v Sloveniji 2009–2010. Ljubljana: Založba ZRC.
- Cao, L. et al. (2012). The application of the spatio-temporal data mining algorithm in maize yield prediction. Mathematical and Computer Modelling, (0). Via http://www.sciencedirect.com/science/article/pii/S0895717711006789 (accessed 04-07-2012).
- Céréghino, R. et al. (2005). Using self-organizing maps to investigate spatial patterns of non-native species. Biological Conservation, 125(4), pp. 459–465.
- Collazos, R. et al. (2006). Análisis espacial del precio de oferta de la vivienda en el área metropolitana de Cochabamba. Revista Latinoamericana de Desarrollo Económico, pp. 633–662.
- Drnovšček, J. (1994). in Prunk et al. (1994). Vodnik po slovenskih vinorodnih okoliših. Ljubljana: Založba Grad.
- European Commission, (2009). Commission Regulation (EC) No 436/2009 of 26 May 2009 laying down detailed rules for the application of Council Regulation (EC) No 479/2008 as regards the vineyard register, compulsory declarations and the gathering of information to monitor the wine market, the documents accompanying consignments of wine products and the wine sector registers to be kept.
- Everingham, Y.L. et al. (2009). Ensemble data mining approaches to forecast regional sugarcane crop production. Agricultural and Forest Meteorology, 149(3–4), pp. 689–696.
- Fayyad, U. et al. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), pp. 37-54.
- Ferraro, D.O. et al. (2009). An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. Field Crops Research, 112(2–3), pp.149–157.
- Foody, G.M. (2003). Geographical weighting as a further refinement to regression modeling: An example focused on the NDVI-rainfall relationship. Remote Sensing of Environment, 88, pp. 283–293.
- Fotheringham, A.S. et al. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environment and Planning A, 30(11), pp.1905–1927.
- Fotheringham, A.S. et al. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships 1st ed., Wiley.
- Giraudel, J.L. & Lek, S. (2001). A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecological Modelling, 146(1-3), pp. 329–339.
- Gouveia, C. et al. (2011). Modelling past and future wine production in the Portuguese Douro Valley. Climate Research, 48(2), pp. 349-362.
- Hand, D. (1998). Data mining: Statistics and more? American Statistician, 52(2), pp. 112–118.
- Hellman, E. (2003). Oregon Viticulture, 1st ed. Oregon: Oregon State University Press.
- Henriques, R. et al. (2012). Exploratory geospatial data analysis using the GeoSOM suite. Computers, Environment and Urban Systems, 36(3), pp. 218–232.
- Hrček, L., & Korošec – Koruza Z. (1996). Sorte in podlage vinske trte: ilustrirani prikaz trsnega izbora za Slovenijo. Ptuj: SVA Veritas.
- Huglin, P. (1986). Biologie et écologie de la vigne. Paris: Editions Payot Lausanne.
- HUT, (2007). Helsinki University of Technology – CIS Laboratory (2007). 'Bibliography of SOM Papers". Via http://www.cis.hut.fi/research/som-bibl/ (accessed 22-04-2012).
- Hutcheson, G. D. (2011). Ordinary Least-Squares Regression. in L. Moutinho & G. D. Hutcheson, The SAGE Dictionary of Quantitative Management Research. pp 224-228.
- Jakša, M. (MAE) (2011). Personal interview. 16-12-2011.

- Kaski, S. & Kohonen, T. (1996). Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. in Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets. World Scientific, pp. 498–507.
- Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), pp. 1464–1480.
- Kuljaj, I. (2005). Trte in vina na Slovenskem. Ljubljana: Magnolija.
- MAE. (2012a). Ministry of Agriculture and Environment of Slovenia data.
- MAE. (2012b). Ministry of Agriculture and Environment of Slovenia public data viewer. Via http://rkg.gov.si/GERK/WebViewer (accessed 07-09-2012).
- Malone, J. et al. (2005). Data mining using rule extraction from Kohonen self-organising maps. Neural Comput. Appl., 15(1), pp. 9–17.
- Martin (2012). Sveti Martin wines web page. Via http://www.vinasvetimartin.si/index.php?page=static&item=476&get_treerot=2 (accessed 07-04-2012).
- Mennis, J. & Guo, D. (2009). Spatial data mining and geographic knowledge discovery-An introduction. Computers Environment and Urban Systems, 33(6), pp. 403–408.
- Miller, J. et al. (2007). Incorporating spatial dependence in predictive vegetation models. Ecological Modelling, 202(3–4), pp. 225–242.
- Ng, R.T. & Han, J. (2002). CLARANS: a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5), pp. 1003–1016.
- Nuske, S. (2012). IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA , SEP 25-30, 2011.
- PARS. (2012). Phytosanitary Service of Slovenia data.
- Perry, E.M. et al. (2009). Spatial variation in tree characteristics and yield in a pear orchard. Precision Agriculture, 11(1), pp. 42–60.
- Pineda Jaimes, N.B. et al. (2010). Exploring the driving forces behind deforestation in the state of Mexico (Mexico) using geographically weighted regression. Applied Geography, 30(4), pp. 576–591.
- Prasad, A.K. et al. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. International Journal of Applied Earth Observation and Geoinformation, 8(1), pp. 26–33.
- R. (2012). A language and environment for statistical computing. Via http://www.r-project.org/ (accessed 16-02-2012 - 13-08-2012).
- R help. (2012). R help page. Via http://r.789695.n4.nabble.com/ (accessed 16-02-2012 - 15-10-2012).
- R kohonen manual. (2012). R kohonen package manual. Via http://cran.r-project.org/web/packages/kohonen/ (accessed 11-04-2012 - 20-08-2012).
- Records, G.W. (2011). Guinness World Records 2012, Guinness World Records.
- Ritter, H. & Kohonen, T. (1989). Self-Organizing Semantic Maps. Biological Cybernetics, 61(4), pp. 241–254.
- Rusjan, D. & Korošec-Koruza Z. (2003). Mikrorajonizacija vinorodnega okoliša Goriška brda. Zbornik Biotehniške fakultete Univerze v Ljubljani. 81(2), pp. 357-367.
- SAM. (2012). SAM Spatial Analysis in Macroeconomy. Via http://www.ecoevol.ufg.br/sam (accessed 18-05-2012).
- SEA. (2012). Slovenian Environmental Agency data. Via http://www.arso.gov.si/vreme/napovedi%20in%20podatki/bilje.html (accessed 18-02-2012).
- Shanmuganathan, S. et al. (2010). Data Mining Techniques for Modelling the Influence of Daily Extreme Weather Conditions on Grapevine, Wine Quality and Perennial Crop Yield. In IEEE, pp. 90–95.
- Stegovec (2012). Stegovec wines vinification description. Via http://klet.cespov-njok.com/category/vinar/stegovec/ (accessed 07-04-2012).
- Stevenson, T. (2005). The Sotheby's Wine Encyclopedia. Dorling Kindersley.
- Tu, J. & Xia, Z. (2008). Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. Science of the Total Environment, 407(1), pp. 358–378.
- Unganai, L. & Kogan, F. (1998). Drought monitoring and corn yield estimation in Southern Africa from AVHRR data RID F-5600-2010. Remote Sensing of Environment, 63(3), pp. 219–232.
- Uradni list, (1999a). Pravilnik o kontroli kakovosti grozdja v času trgatve, 1999. Uradni list Republike Slovenije, Ur.l. RS, št. 68/1999; Ur.l. RS, št. 79/2000, 69/2001.

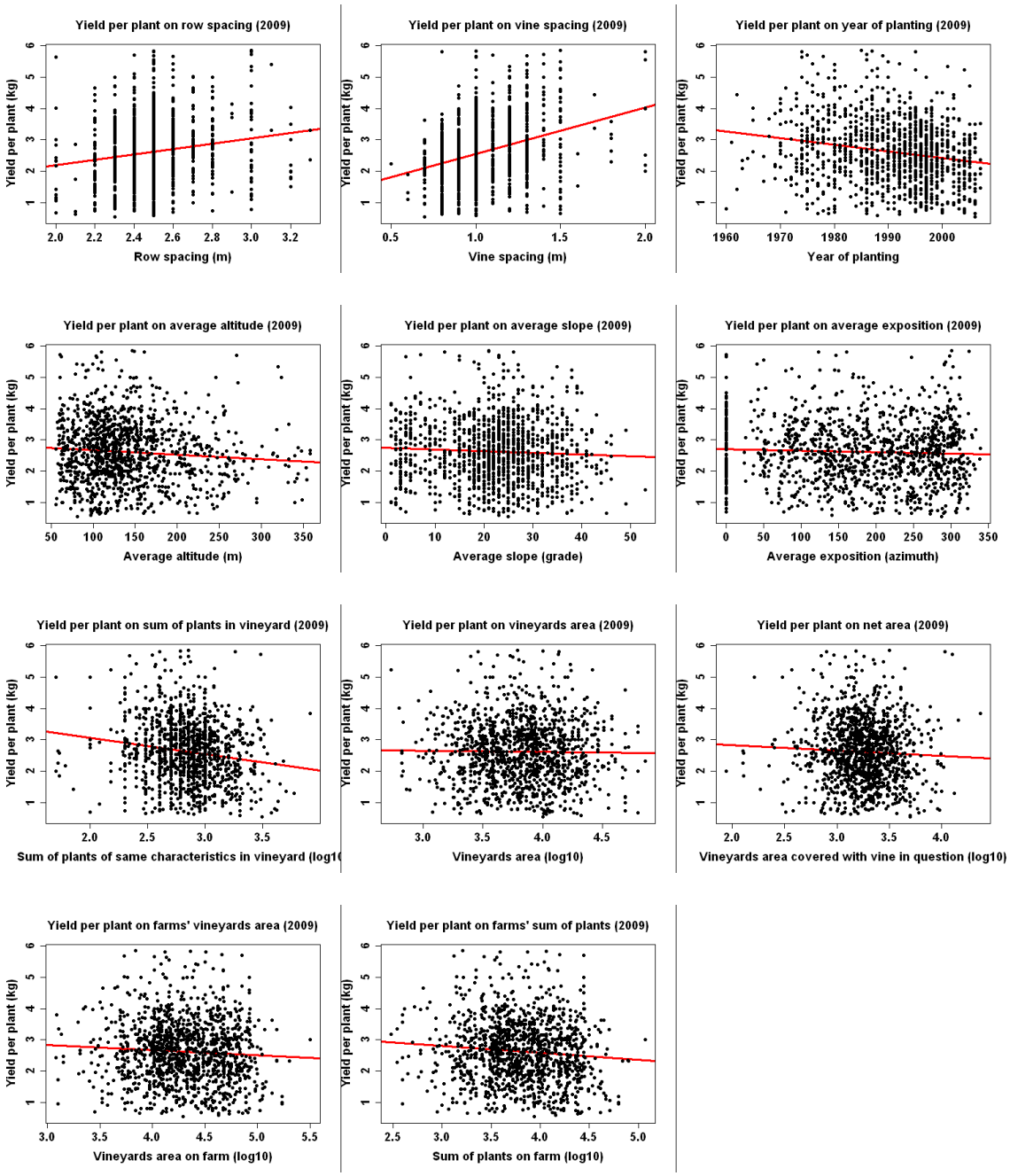
- Uradni list, (1999b). Pravilnik o registru pridelovalcev grozdja in vina in katastru vinogradov. 2007. Uradni list Republike Slovenije, Ur.l. RS, št. 44/1999; Ur.l. RS, št. 79/2000, 5/2004, 16/2007.
- Uradni list, (2004). Pravilnik o pogojih, ki jih mora izpolnjevati grozdje za predelavo v vino, o dovoljenih tehnoloških postopkih in enoloških sredstvih za pridelavo vina in o pogojih glede kakovosti vina, mošta in drugih proizvodov v prometu. 2004. Uradni list Republike Slovenije, Ur.l. RS, št. 43/2004; Ur.l. RS, št. 127/2004, 112/2005.
- Uradni list, (2003). Pravilnik o razdelitvi vinogradniškega območja v Republiki Sloveniji, absolutnih vinogradniških legah in o dovoljenih ter priporočenih sortah vinske. 2003. Uradni list Republike Slovenije, Ur.l. RS, št. 69/2003; Ur.l. RS, št. 31/2004, 117/2004, 49/2007.
- Uradni list, (2006). Pravilnik o registru kmetijskih gospodarstev. 2006. Uradni list Republike Slovenije, Ur.l. RS, št. 121/2006; Ur.l. RS, št. 124/2007, 45/2008-ZKme-1, 122/2008.
- Venables W. N. et al. (2012). An introduction to R. Via http://cran.rproject.org/doc/manuals/R-intro.pdf (accessed 16-02-2012).
- Vesanto, J. (1997). Data Mining Techniques Based on the Self-Organizing Map (Msc thesis), Helsinki University of Technology, Department of Engineering Physics and Mathematics. Via http://www.cis.hut.fi/projects/ide/publications/html/mastersJV97/ (accessed 22-04-2012).
- Vesanto, J. et al. (2000). SOM Toolbox for Matlab 5, Report A57. Via at: http://www.cis.hut.fi/projects/somtoolbox/ (accessed 22-04-2012).
- Vesanto, J. (2002). Data Exploration Process Based on the Self-Organizing Map (Phd thesis), Helsinki University of Technology, Department of Computer Science and Engineering. Via http://lib.tkk.fi/Diss/2002/isbn9512258978/ (accessed 22-04-2012).
- Vršič, S. & Lešnik, M. (2001). Vinogradništvo. Ljubljana: Kmečki glas.
- Wang, Q. (2005). Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. Global Ecology and Biogeography, 14(4), pp. 379–393.
- Wehrens, R. & Buydens, L.M.C. (2007). Self-and super-organizing maps in R: the Kohonen package. Journal Of Statistical Software, 21(5), pp. 1-19.
- WOK. (2012). Web of knowledge. Via http://wok.mimas.ac.uk/ (accessed 25-02-2012).
- Wolpert, J.A. & Vilas, E.P. (1992). Estimating Vineyard Yields: Introduction to a Simple, Two-Step Method. American Journal of Enology and Viticulture, 43(4), pp. 384 -388.
- Yao, X. (2003), Research issues in spatio-temporal data mining, a white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, Nov. 18-20.
- Zupančič, P. (1995), Klimatografija Slovenije 1961 -1990. Padavine. Ljubljana: Ministrstvo za okolje in prostor, Hidrometeorološki zavod RS.

APPENDICES

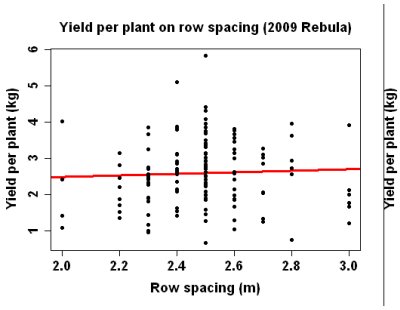
Appendix A:

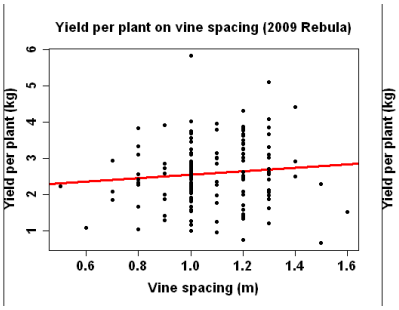
Scatter plots of 2009 data: the correlations between all pairs of a dependent variable and explanatory variables.

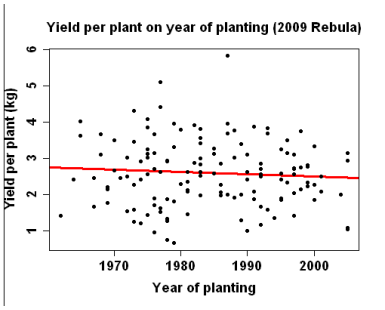
All data:

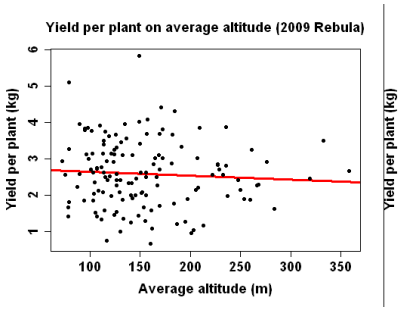


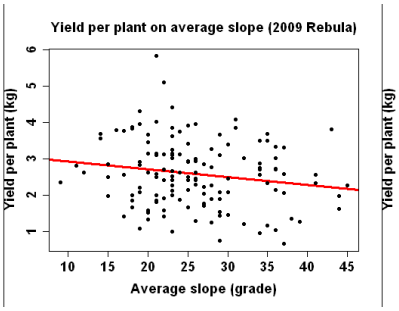
Rebula data:

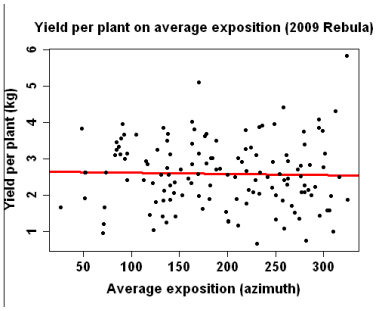

Yield per plant on row spacing (2009 Rebula)
Yield per plant (kg)
Row spacing (m)

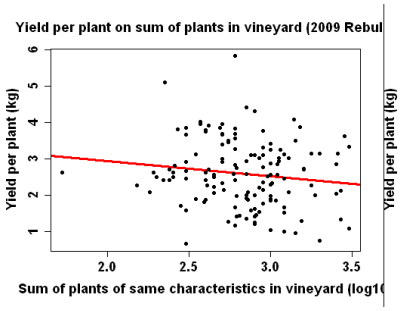

Yield per plant on vine spacing (2009 Rebula)
Yield per plant (kg)
Vine spacing (m)

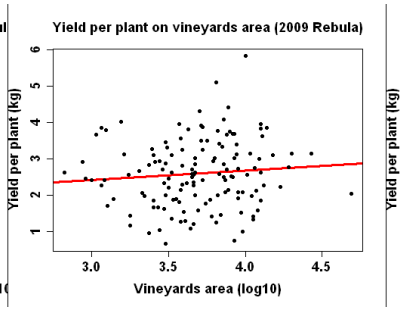

Yield per plant on year of planting (2009 Rebula)
Yield per plant (kg)
Year of planting

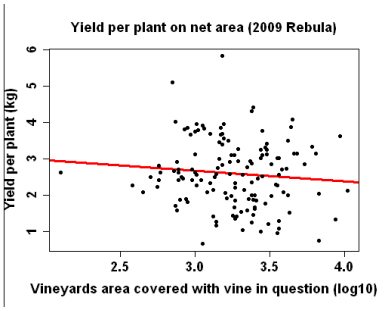

Yield per plant on average altitude (2009 Rebula)
Yield per plant (kg)
Average altitude (m)

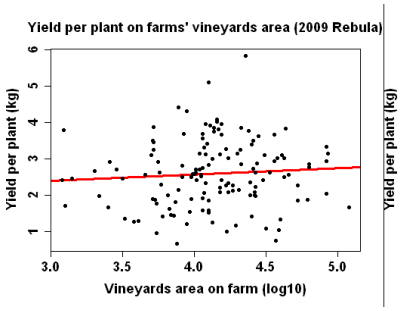

Yield per plant on average slope (2009 Rebula)
Yield per plant (kg)
Average slope (grade)

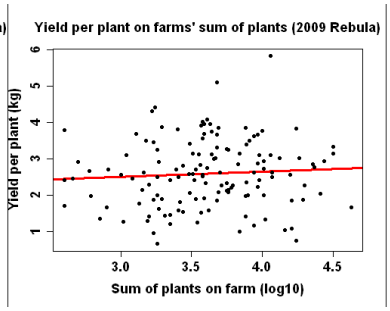

Yield per plant on average exposition (2009 Rebula)
Yield per plant (kg)
Average exposition (azimuth)


Yield per plant on sum of plants in vineyard (2009 Rebul
Yield per plant (kg)
Sum of plants of same characteristics in vineyard (log1


Yield per plant on vineyards area (2009 Rebula)
Yield per plant (kg)
Vineyards area (log10)


Yield per plant on net area (2009 Rebula)
Yield per plant (kg)
Vineyards area covered with vine in question (log10)


Yield per plant on farms' vineyards area (2009 Rebula)
Yield per plant (kg)
Vineyards area on farm (log10)

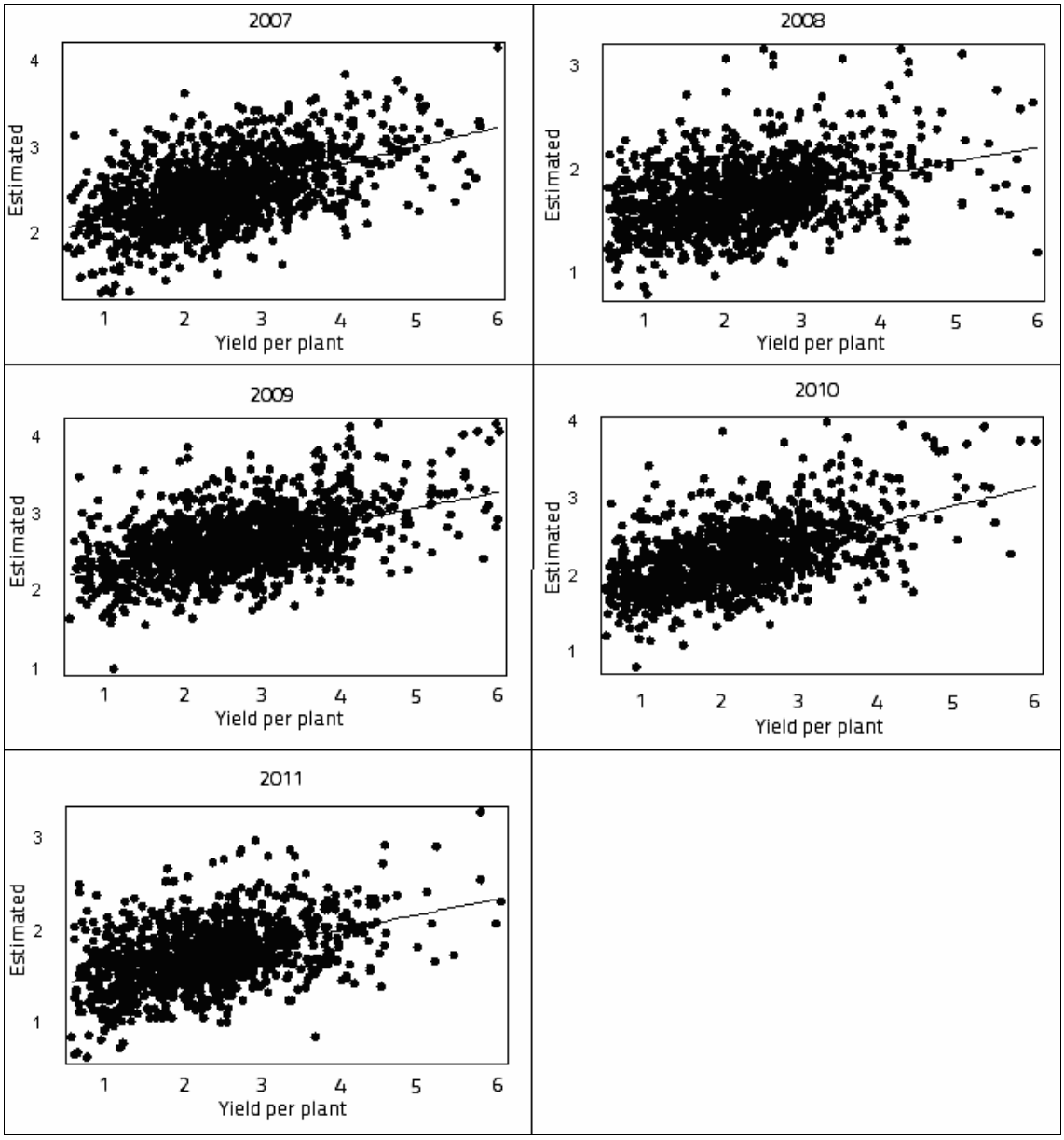

Yield per plant on farms' sum of plants (2009 Rebula)
Yield per plant (kg)
Sum of plants on farm (log10)

Appendix B:

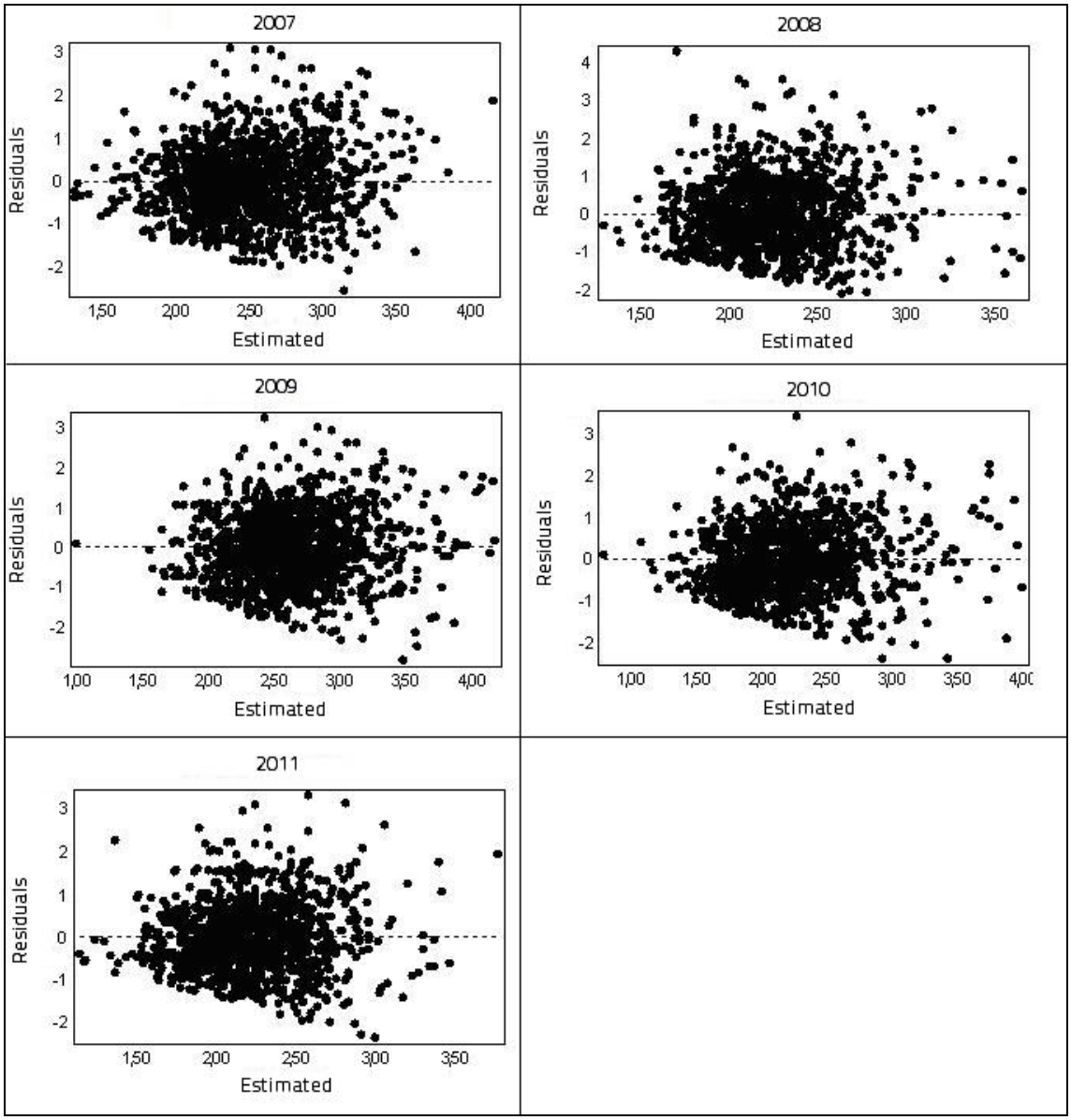
Selected GWR outputs (for all data) including:

1. plots of GWR estimation on actual yield per plant,
2. plots of residuals to estimated values,
3. local R^2 histograms,
4. residuals histograms,
5. maps of residuals.

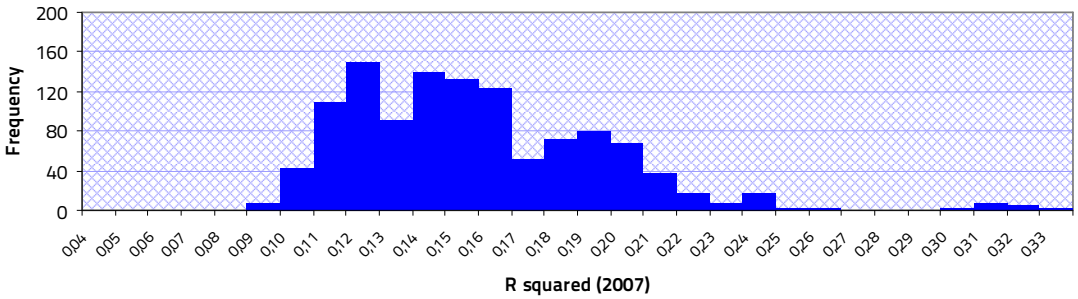
1. Plots of GWR estimation on actual yield per plant for 2007 – 2011 respectively:

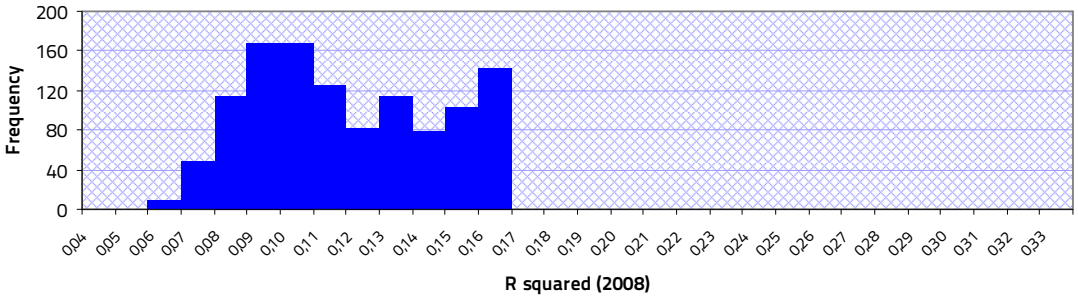


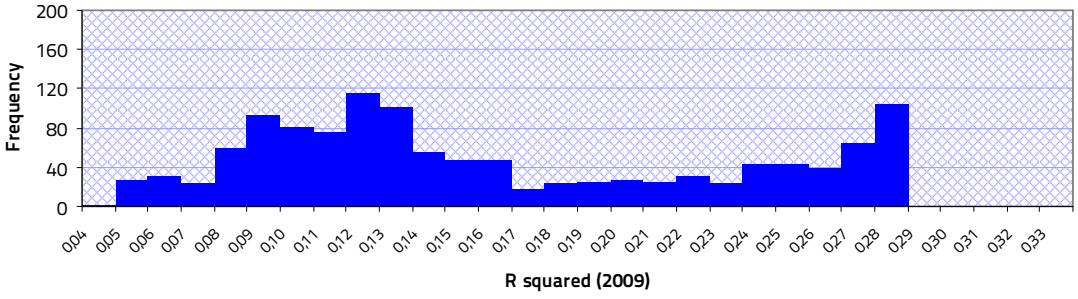
2. Plot of GWR residuals to estimated values for 2007 – 2011 respectively:

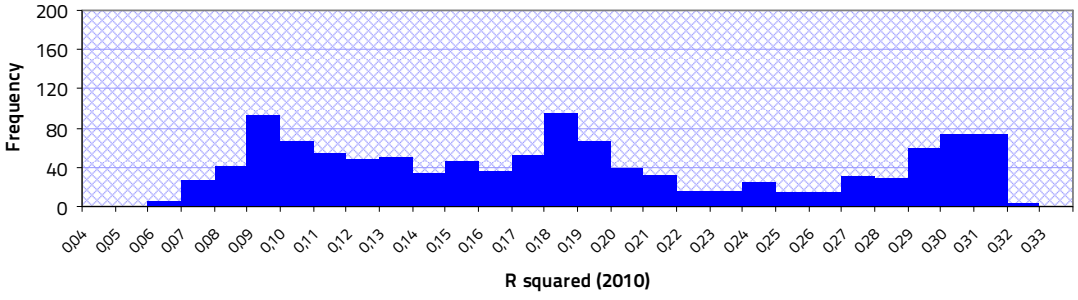


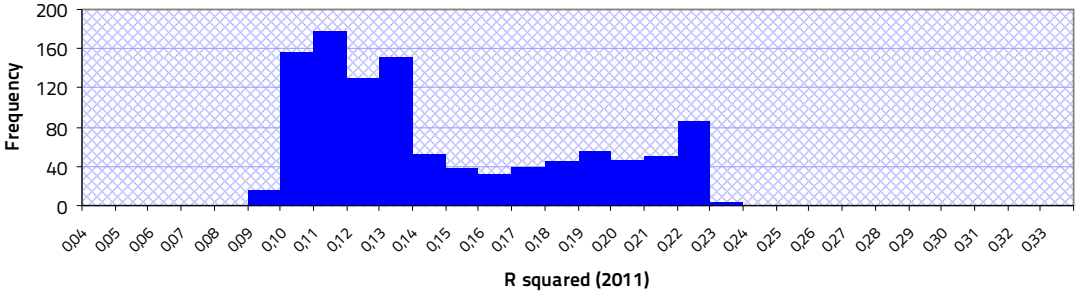
3. Local R^2 histograms for 2007 – 2011 respectively:



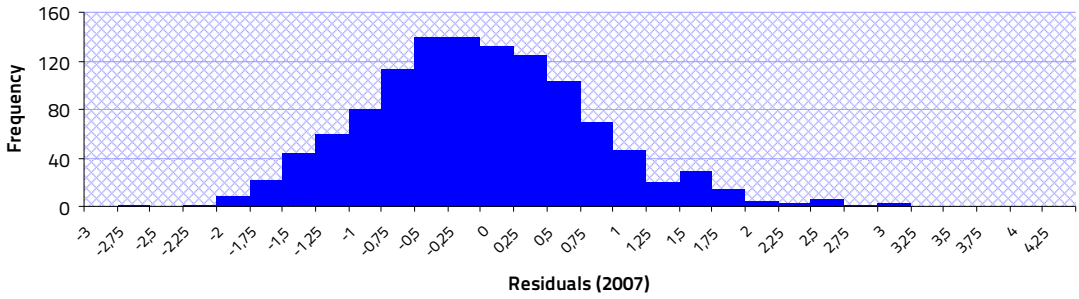


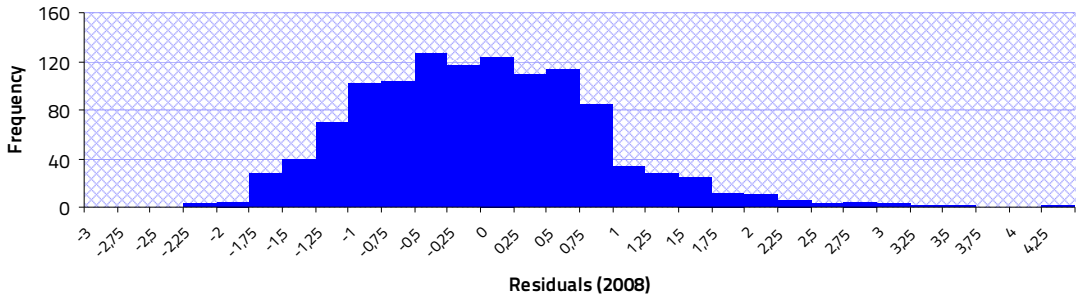


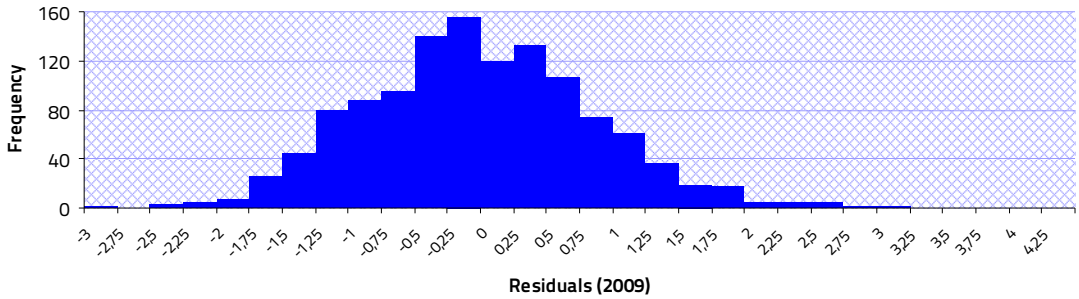


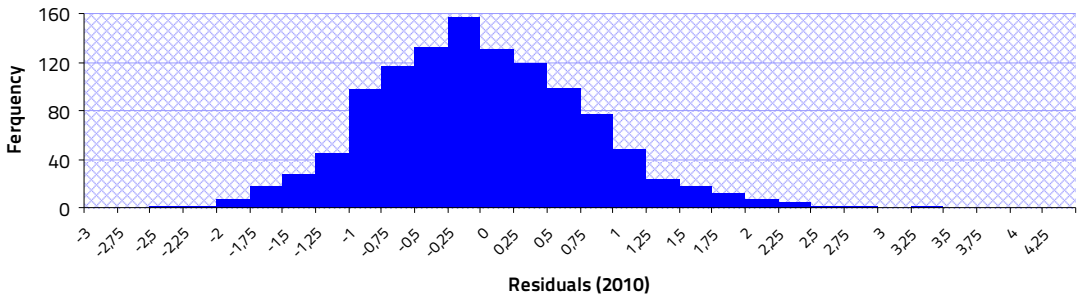


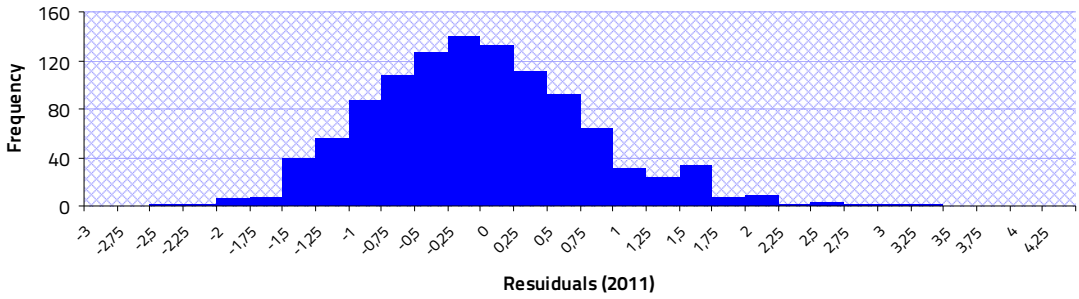
4. Residuals histograms for 2007 – 2011 respectively:



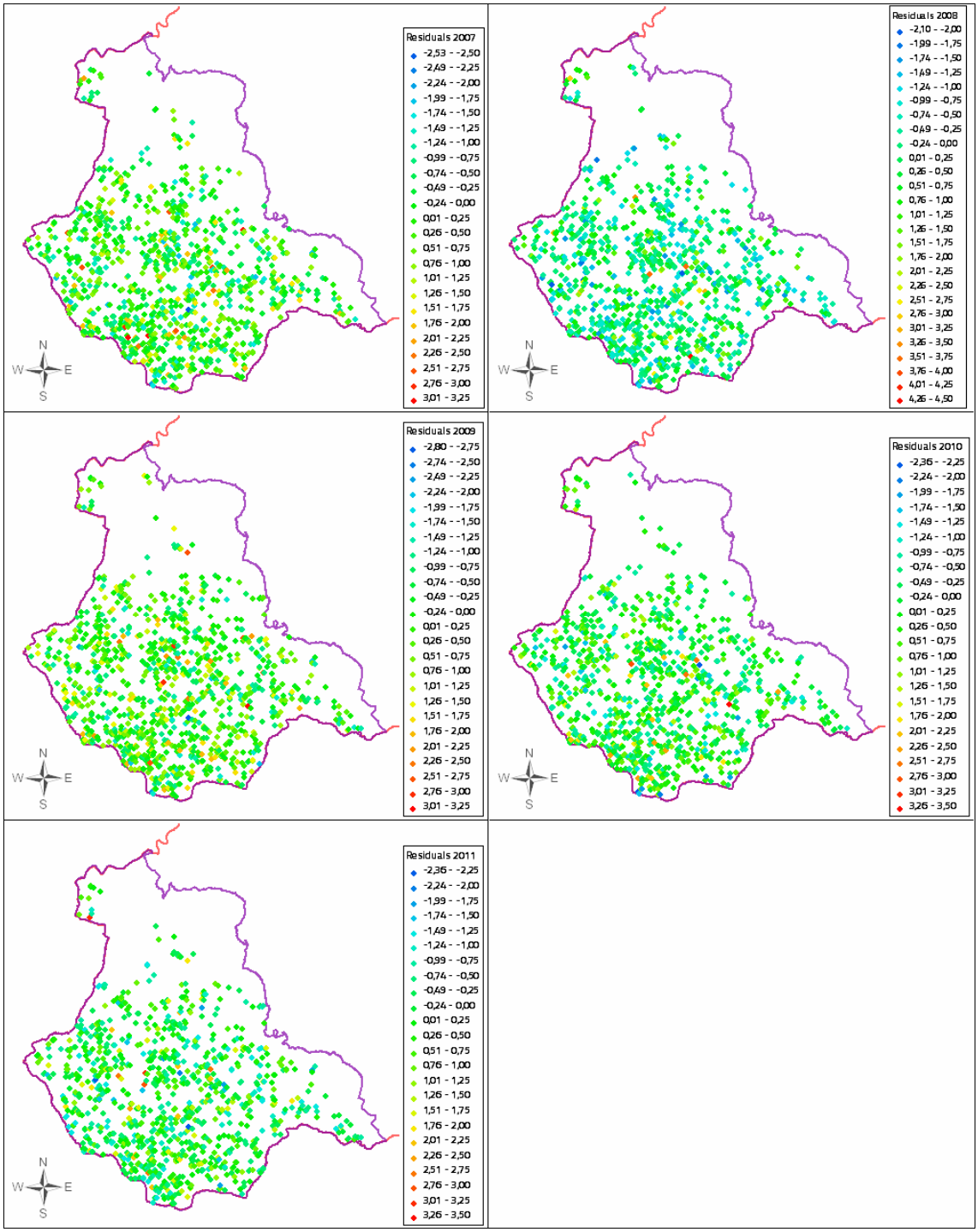








5. Maps of residuals for 2007 – 2011 respectively:

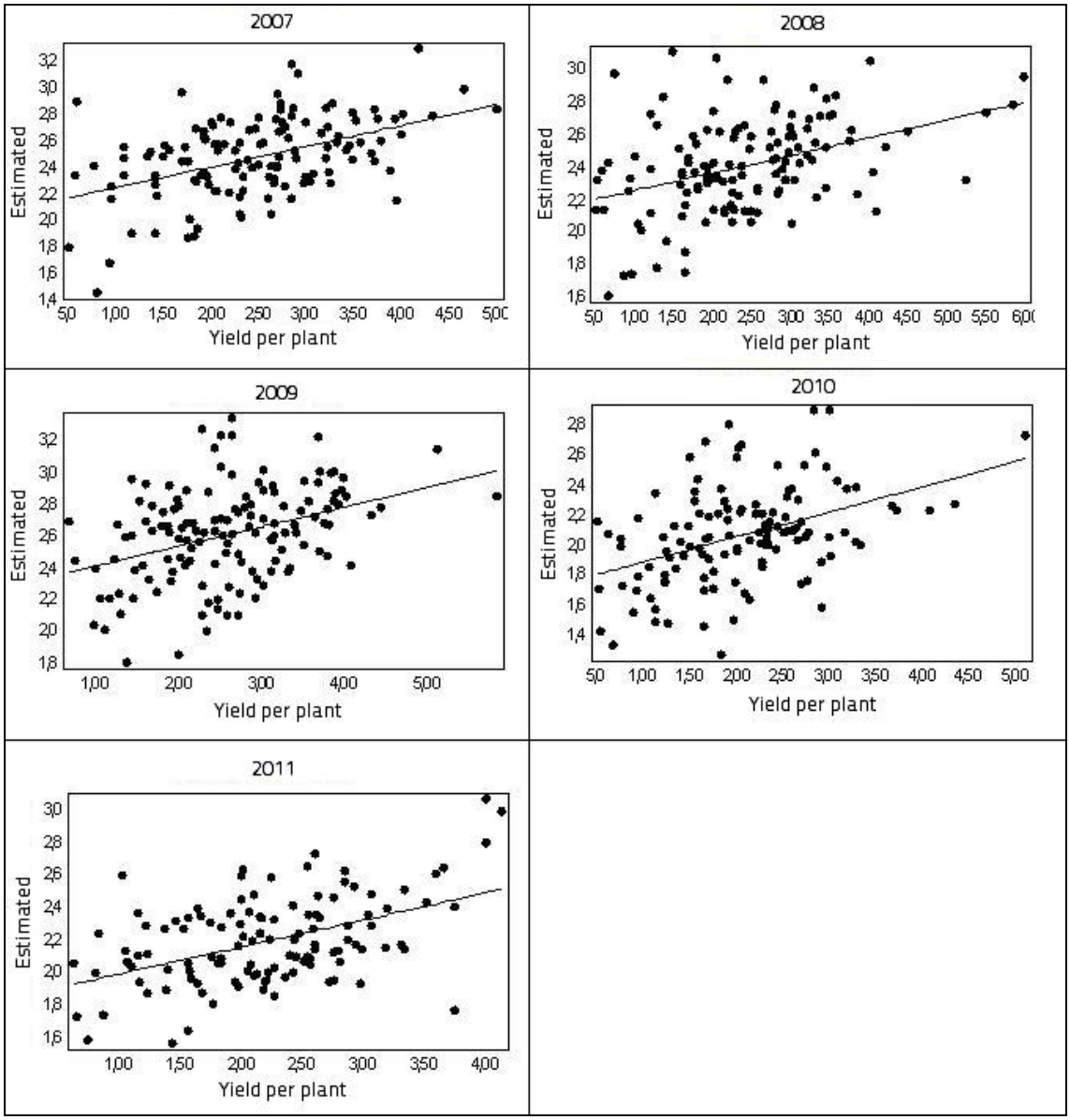


Appendix C:

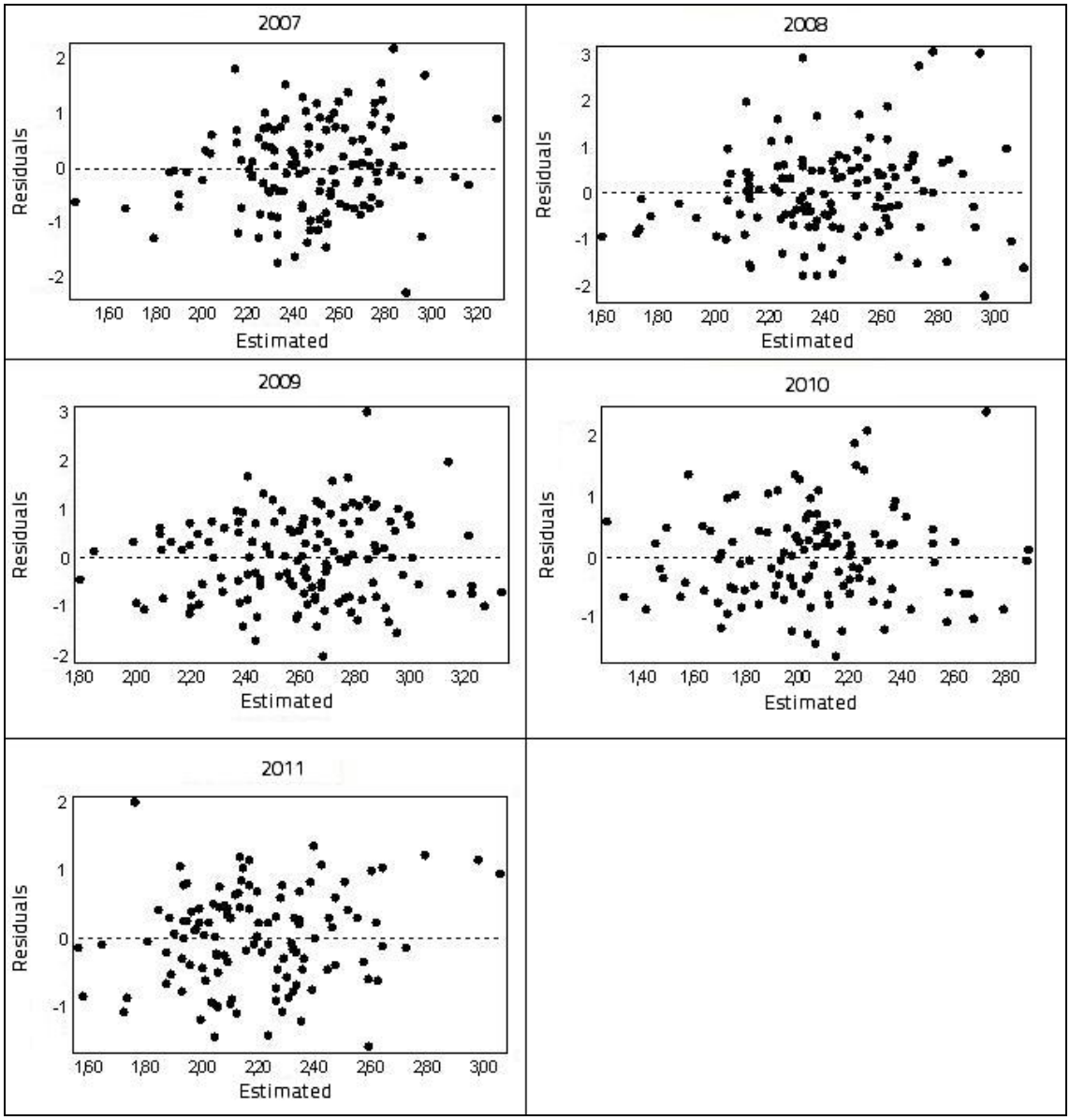
Selected GWR outputs (for Rebula) including:

1. plots of GWR estimation on actual yield per plant,
2. plots of residuals to estimated values,
3. local R^2 histograms,
4. residuals histograms,
5. maps of residuals.

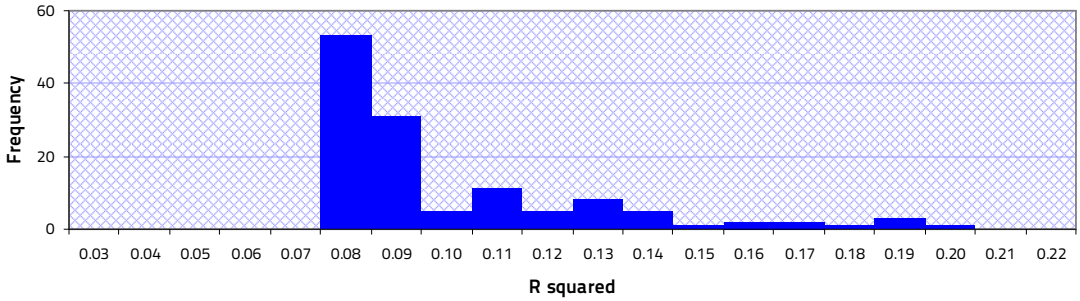
1. Plot of GWR estimation on actual yield per plant for 2007 – 2011 respectively:

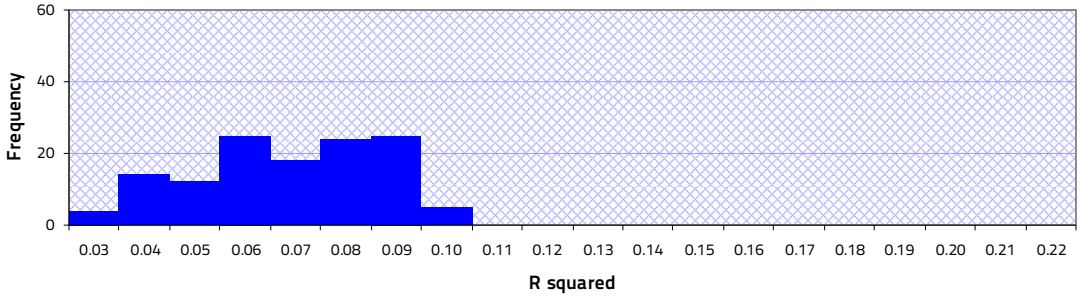


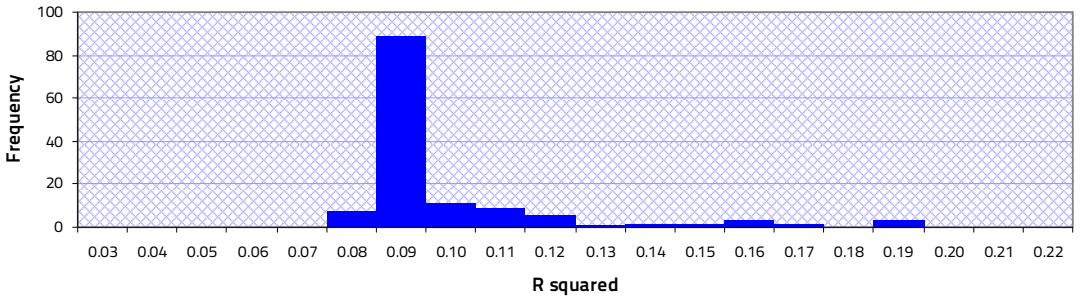
2. Plot of GWR residuals to estimated values for 2007 – 2011 respectively:

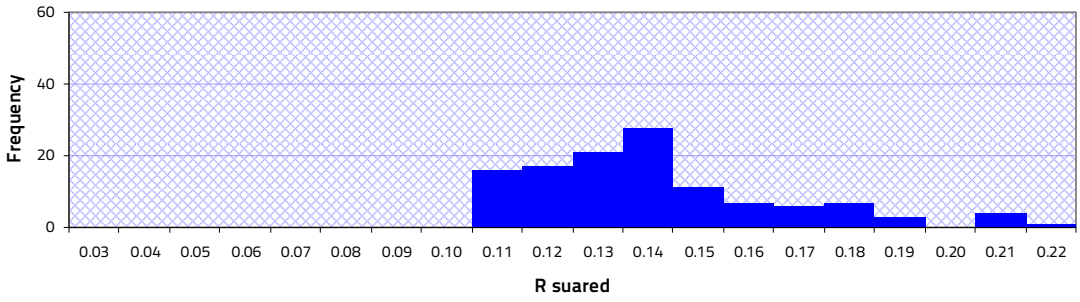


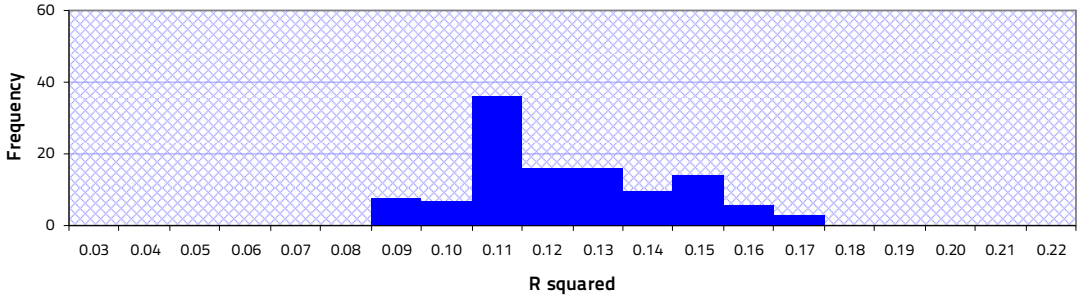
3. Local R^2 histograms for 2007 – 2011 respectively:



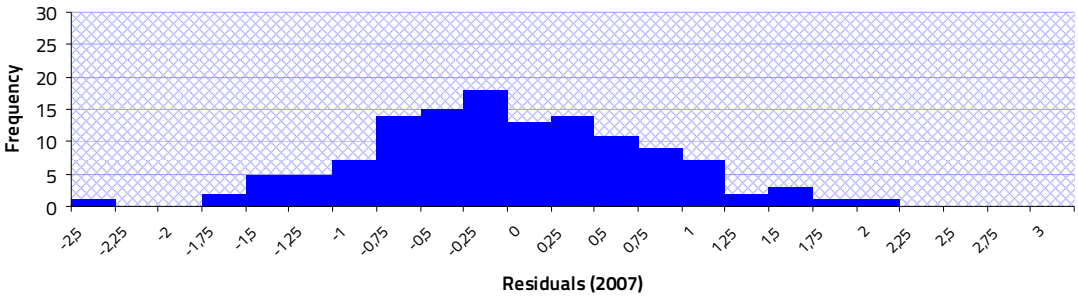


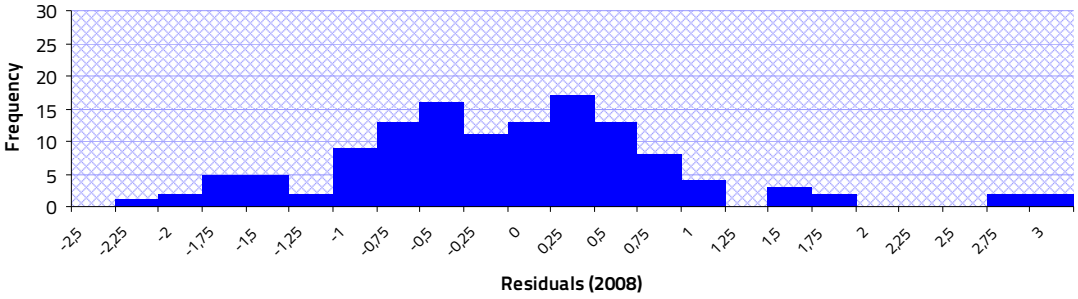


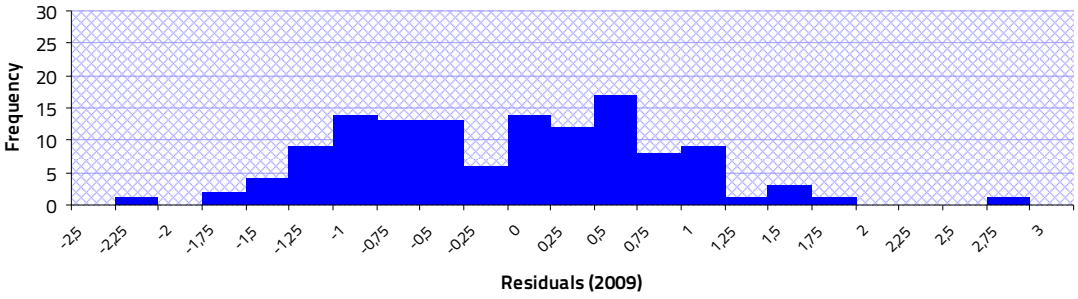


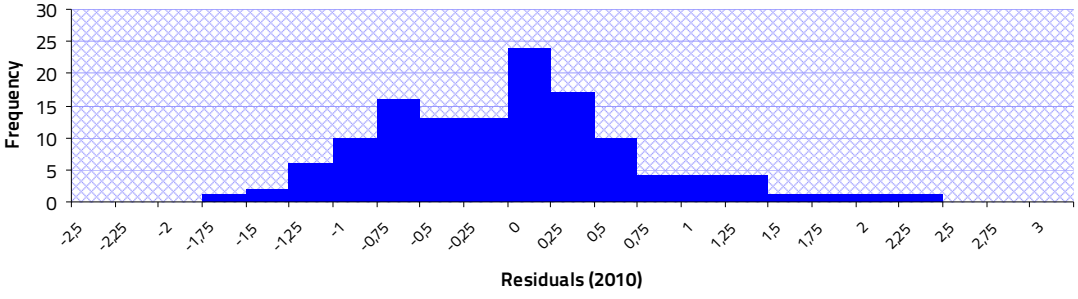


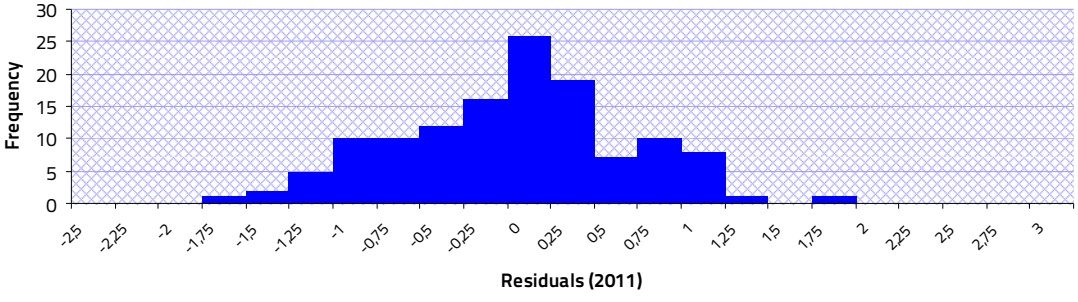
4. Residuals histograms for 2007 – 2011 respectively:



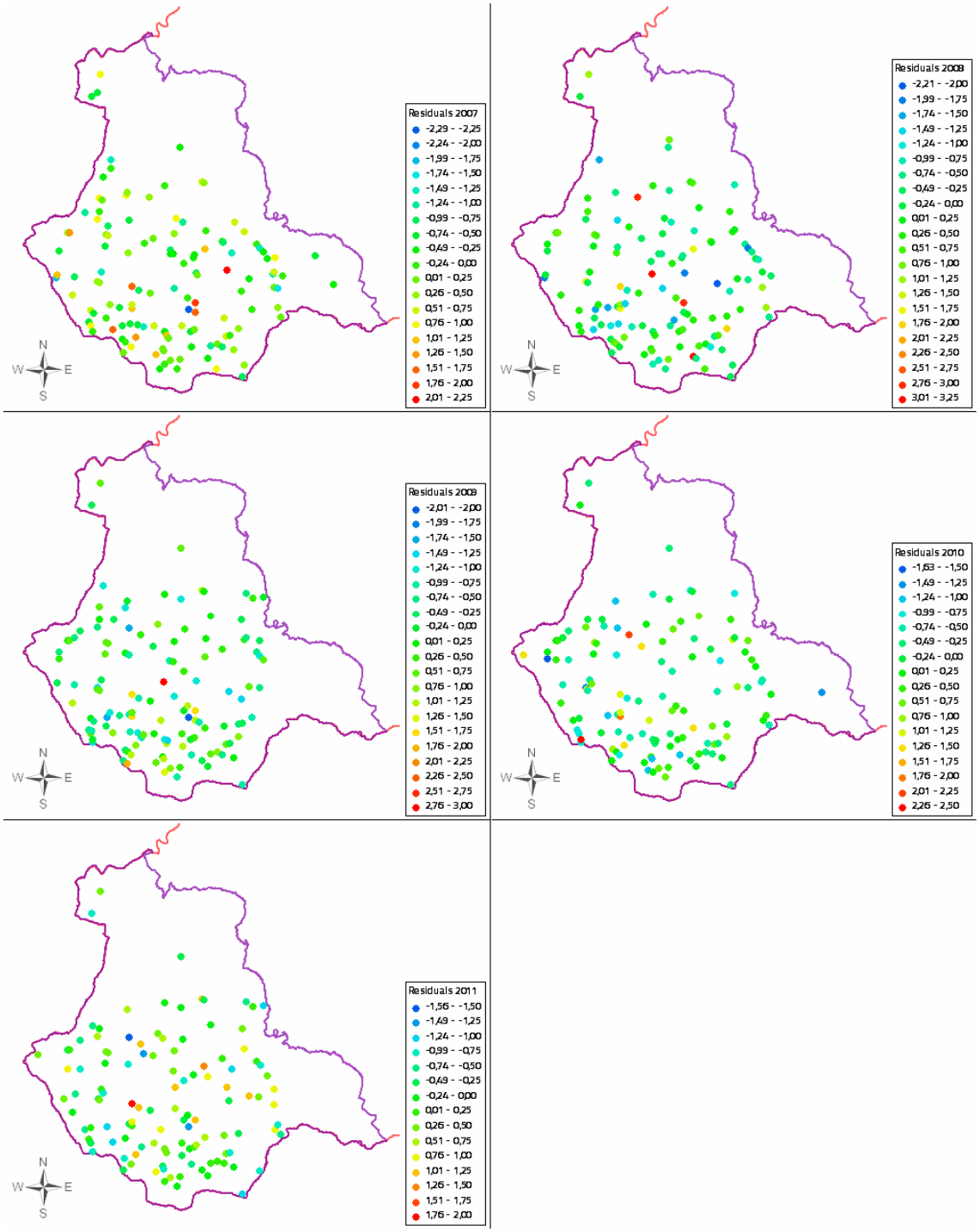








5. Maps of residuals for 2007 – 2011 respectively:

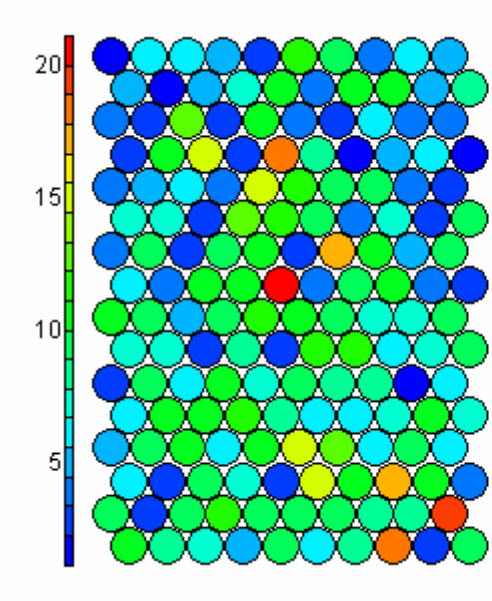


Appendix D:

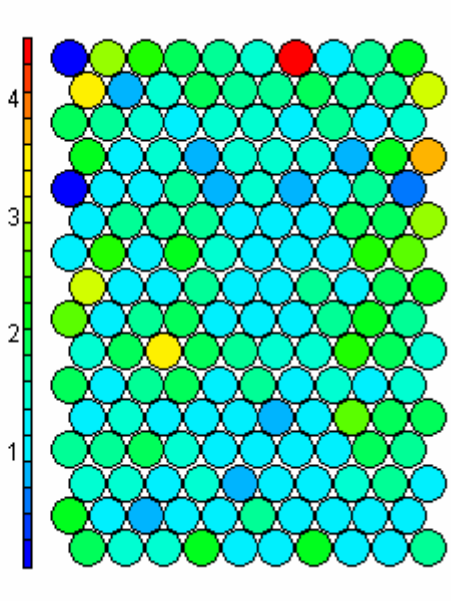
Selected SOM outputs (for all data) including:

1. Count of occurrences plot,
2. Mapping quality plot,
3. Distance change plot.

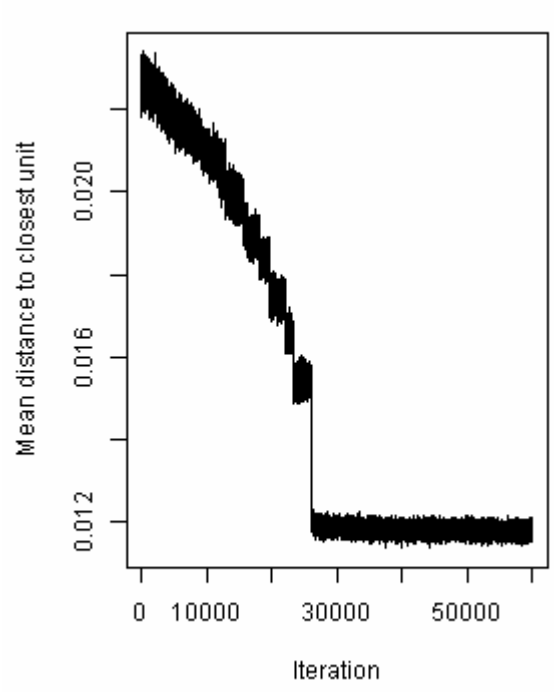
1. Count of occurrences plot:



2. Mapping quality plot:



3. Distance change plot:

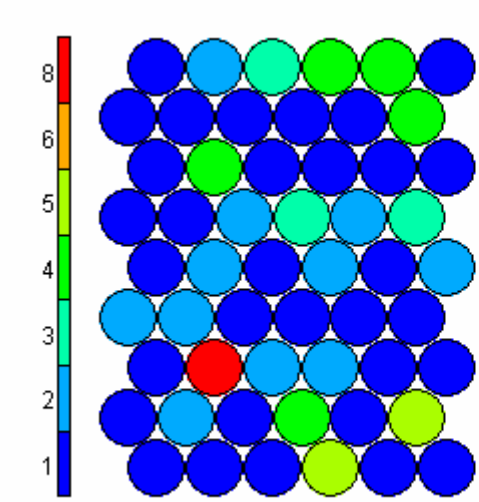


Appendix E:

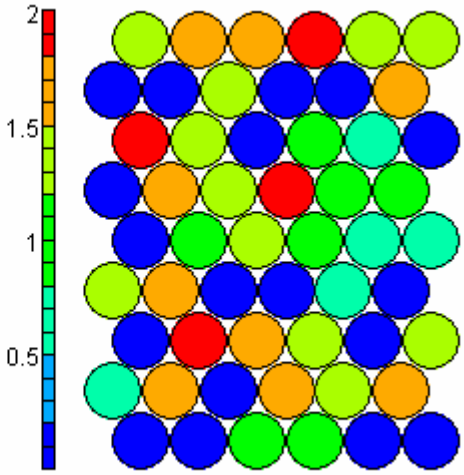
Selected SOM outputs (for Rebula data) including:

1. Count of occurrences plot,
2. Mapping quality plot,
3. Distance change plot.

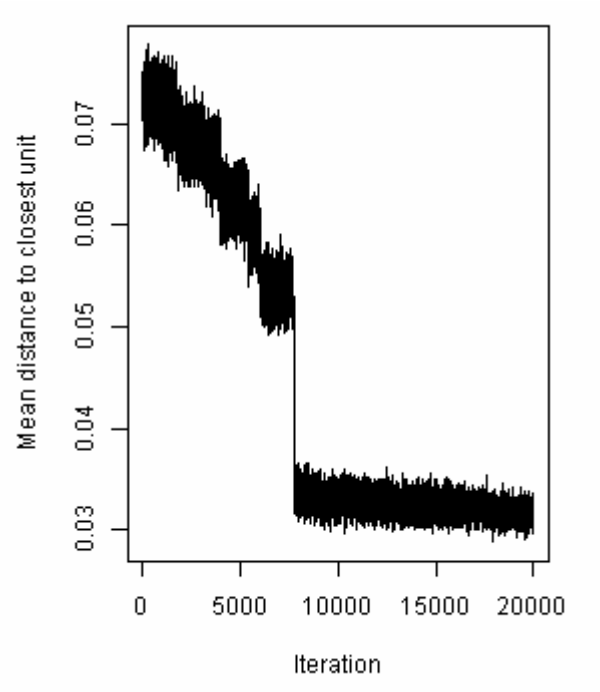
1. Count of occurrences plot:



2. Mapping quality plot:



3. Distance change plot:



Appendix F:

R scripts including:

1. Data preparation script (2009 all data example),
2. VIF review script,
3. SOM training and graphics output script (2009 all data example),
4. SOM mapping script.

1. Data preparation script (2009 all data example):

```
#1. loading packages and data
library(foreign)
library(plyr)
GERK = read.dbf("./GERK_2009.dbf")
TNS = read.dbf("./TNS_2009.dbf")
PRI = read.dbf("./PRI_2009.dbf")

#2. selecting of GERKs, vineyards and yield records within Goriška Brda wine district (PKE attributed in ARCGis
1=within Brda, 0=outside Brda):
GERK_Gb<-GERK[GERK$PKE > 1,]
GERK_xGb<-GERK[GERK$PKE == 0,]
GERK_ok <- GERK_Gb
GERK_ok <- GERK_ok [with(GERK_ok, order(KMG_MID, GERK_PID)),]
GERK_xok <- GERK_xGb
GERK_xok <- GERK_xok [with(GERK_xok, order(KMG_MID, GERK_PID)),]
GERK_xok_TNS <- merge (GERK_xok, TNS)
TNS_si <- TNS[!TNS$GERK_PID %in% GERK_xok_TNS$GERK_PID,]
GERK_xok_TNS$link = paste(GERK_xok_TNS$KMG_MID, GERK_xok_TNS$SORTA_ID, sep='')
PRI$link = paste(PRI$KMG_MID, PRI$SORTA_ID, sep='')
PRI_si <- PRI[!PRI$link %in% GERK_xok_TNS$link,]
PRI_xsi <- PRI[PRI$link %in% GERK_xok_TNS$link,]

#3 selecting appropriate vineyards records and removing of (year of planting) outliers:
TNS_si_v2 <- subset (TNS_si, LETO_SAJEN < 2008 & LETO_SAJEN > 1959)
TNS_si_v2$SUM_SADIK <- rep(1,nrow(TNS_si_v2))
TNS_si_v3 <- ddply(TNS_si_v2, c("KMG_MID","GERK_PID","SORTA_ID","LETO_SAJEN","SR_MED_VRS","SR_V_VRSTI","GOJ_OBL_ID"),
summarize, SUM_SADIK=sum(STEVILO_SA))
TNS_si_v4 <- TNS_si_v3
TNS_si_v4$link = paste(TNS_si_v4$KMG_MID, TNS_si_v4$SORTA_ID, sep='')
TNS_si_v4$multiple <- rep(1,nrow(TNS_si_v4))
TNS_si_v5 <- ddply(TNS_si_v4, c("link"), summarize, multiple=sum(multiple))
TNS_si_v6<-TNS_si_v5[TNS_si_v5$multiple == 1,]
TNS_si_v7 <- merge (TNS_si_v6, TNS_si_v4)
TNS_ok <- TNS_si_v7

#4 selecting appropriate yield records:
PRI_si$SUM_KOLICINE <- rep(1,nrow(PRI_si))
PRI_si_v1 <- ddply(PRI_si, c("KMG_MID","SORTA_ID","ID_PRID_EN","DATUM_TR"), summarize, SUM_KOLICINE=sum(KOLICINA))
PRI_si_v2 <- PRI_si_v1
PRI_si_v2$link = paste(PRI_si_v2$KMG_MID, PRI_si_v2$SORTA_ID, sep='')
PRI_si_v2$multiple <- rep(1,nrow(PRI_si_v2))
PRI_si_v3 <- ddply(PRI_si_v2, c("link"), summarize, multiple=sum(multiple))
PRI_si_v4<-PRI_si_v3[PRI_si_v3$multiple == 1,]
PRI_si_v5 <- merge (PRI_si_v4, PRI_si_v2)
PRI_sok <- PRI_si_v5
PRI_mb<-PRI_sok[PRI_sok$SORTA_ID==41,]
colnames(PRI_mb)[2] = "mb"
PRI_mr<-PRI_sok[PRI_sok$SORTA_ID==42,]
colnames(PRI_mr)[2] = "mr"
PRI_join_del <- merge (PRI_sok, PRI_mb, all=TRUE)
PRI_v2 <- merge (PRI_join_del, PRI_mr, all=TRUE)
PRI_v2$mbmr_izbris <- rep(1,nrow(PRI_v2))
PRI_v2$mbmr_izbris <- ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 3, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 6, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 13, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 25, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 29, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 31, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 32, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 48, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 7, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 18, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 21, 1,
ifelse(PRI_v2$mb == 41  &  PRI_v2$SORTA_ID == 27, 1,
ifelse(PRI_v2$SORTA_ID == 41, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 4, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 5, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 14, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 16, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 26, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 1, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 8, 1,
ifelse(PRI_v2$mr == 42  &  PRI_v2$SORTA_ID == 47, 1,
ifelse(PRI_v2$SORTA_ID == 42, 1,
```

```
ifelse(PRI_v2$mb == NA & PRI_v2$mr == NA, 0, 9)))))))))))))))))))))))))
PRI_v2$mbmr_izbris[ is.na(PRI_v2$mbmr_izbris) ] <- 0
PRI_ok<-PRI_v2[PRI_v2$mbmr_izbris == 0,]

#5 joininig (partial), selecting variables
DATA <- merge(merge(PRI_ok, TNS_ok,),GERK_ok,)
DATA$X <- DATA$X + DATA$SORTA_ID
DATA$Y <- DATA$Y + DATA$SORTA_ID
DATA_v1 <- DATA [,c("KMG_MID","GERK_PID","AREA","SORTA_ID","SUM_SADIK","SUM_KOLICINE","LETO_SAJEN","EXP_AVG","Z_AVG",
"NAGIB_AVG","SR_MED_VRS","SR_V_VRSTI","Y", "X")]

#6 calculating dependent variable and new explanatory variables, transforming (log10) certain variables
DATA_v1$KG_na_SADIKO = DATA_v1$SUM_KOLICINE / DATA_v1$SUM_SADIK
DATA_v1$neto_area = DATA_v1$SUM_SADIK * DATA_v1$SR_MED_VRS * DATA_v1$SR_V_VRSTI
GERK_soc <- ddply(GERK, c("KMG_MID"), summarize, sum_vin=sum(AREA))
TNS_soc <- ddply(TNS, c("KMG_MID"), summarize, sum_sad=sum(STEVILO_SA))
DATA_soc <- merge(GERK_soc,TNS_soc)
DATA_soc$sum_vin_l <- log10(DATA_soc$sum_vin)
DATA_soc$sum_sad_l <- log10(DATA_soc$sum_sad)
DATA_v1$SUM_SADIK_l <- log10(DATA_v1$SUM_SADIK)
DATA_v1$SUM_KOLICINE_l <- log10(DATA_v1$SUM_KOLICINE)
DATA_v1$AREA_l <- log10(DATA_v1$AREA)
DATA_v1$neto_area_l <- log10(DATA_v1$neto_area)

#7 final joining, removing outliers (dependent variable) and selecting variables
DATA_v1_soc <- merge(DATA_v1,DATA_soc,)
DATA_ok_stat <- subset (DATA_v1_soc, KG_na_SADIKO > 0.5 & KG_na_SADIKO < 6)
DATA_ok <- DATA_ok_stat [,c("KMG_MID","GERK_PID","SORTA_ID","SUM_SADIK_l","LETO_SAJEN","SR_MED_VRS","SR_V_VRSTI",
"Z_AVG","NAGIB_AVG","EXP_AVG","AREA_l","neto_area_l","SUM_KOLICINE_l","KG_na_SADIKO","sum_vin_l","sum_sad_l","Y", "X")]

#8 renaming variables
names(DATA_ok)[1]<-paste("FARM_ID")
names(DATA_ok)[2]<-paste("PARCEL_ID")
names(DATA_ok)[3]<-paste("VARI_ID")
names(DATA_ok)[4]<-paste("PLANT-SUM")
names(DATA_ok)[5]<-paste("Y_PLANTING")
names(DATA_ok)[6]<-paste("R_SPACING")
names(DATA_ok)[7]<-paste("V_SPACING")
names(DATA_ok)[8]<-paste("Z_AVG")
names(DATA_ok)[9]<-paste("SLOPE_AVG")
names(DATA_ok)[10]<-paste("EXP_AVG")
names(DATA_ok)[11]<-paste("PARC-AREA")
names(DATA_ok)[12]<-paste("NET-AREA")
names(DATA_ok)[13]<-paste("VARI-YI-SUM")
names(DATA_ok)[14]<-paste("YI_PER_PLANT")
names(DATA_ok)[15]<-paste("F-AREA-SUM")
names(DATA_ok)[16]<-paste("F-PLANT-SUM")
names(DATA_ok)[17]<-paste("Y")
names(DATA_ok)[18]<-paste("X")

#9 exporting data
DATA_ok_25 <- subset (DATA_ok, VARI_ID == 25)
write.dbf (DATA_ok_25, "DATA_rebula.dbf", factor2char = TRUE, max_nchar=254)
write.dbf (DATA_ok, "DATA_all.dbf", factor2char = TRUE, max_nchar=254)
```

2. VIF review script:

```
#1 loading packages and data:
library(foreign)
DATA_ok_07 = read.dbf ("./DATA_07.dbf")

#1 deriving VIF critical values by iteration until needed
DATA_ok <- DATA_ok_07
model1 = lm(YI_PER_PLA~PLANT.SUM+Y_PLANTING+R_SPACING+V_SPACING+Z_AVG+SLOPE_AVG+EXP_AVG+PARC.AREA+NET.AREA+F.AREA.SUM+
F.PLANT.SU, data=DATA_ok)
vif (model1)
model2 = lm (YI_PER_PLA~Y_PLANTING+R_SPACING+V_SPACING+Z_AVG+SLOPE_AVG+EXP_AVG+PARC.AREA+NET.AREA+F.AREA.SUM+F.PLANT.SU,
data=DATA_ok)
vif (model2)
model3 = lm (YI_PER_PLA~Y_PLANTING+R_SPACING+V_SPACING+Z_AVG+SLOPE_AVG+EXP_AVG+PARC.AREA+NET.AREA+F.AREA.SUM, data=
DATA_ok)
vif (model3)
```

3. SOM training script and graphics output script (2009 all data example):

```
#1. loading packages and data:
library("kohonen")
library (foreign)
library (plyr)
DATA_ok_09 = read.dbf ("./DATA_09.dbf")

#2. scaling data
DATA_ok_09 <- DATA_ok_09 [,c("NET.AREA","Y_PLANTING","B_R_DIST","W_R_DIST","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM",
"PARC.AREA")]
DATA_oks_09 <- scale (DATA_ok_09)

#3. training SOM
set.seed(7)
Brda09.som <- som (data = DATA_oks_09,
grid = somgrid(10, 16, "hexagonal"),
rlen = 60000,
alpha = c(0.05, 0.01)),
radius = 0.67)

#4. setting palette and drawing plots
coolBlueHotRed <- function(n, alpha = 1) {rainbow(n, end=4/6, alpha=alpha)[n:1]}
plot(Brda09.som, type = "dist.neighbours", main = "U-matrix, all, 2009", palette.name = coolBlueHotRed)
plot(Brda09.som, type = "quality", main = "Mapping quality, all, 2009", palette.name = coolBlueHotRed)
plot(Brda09.som, type = "counts", main = "Count of occurences, all, 2009", palette.name = coolBlueHotRed)
plot(Brda09.som, type = "changes")
plot(Brda09.som, type = "property", property = Brda09.som$codes[,1], main = colnames(Brda09.som$codes)[1])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,2], main = colnames(Brda09.som$codes)[2])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,3], main = colnames(Brda09.som$codes)[3])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,4], main = colnames(Brda09.som$codes)[4])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,5], main = colnames(Brda09.som$codes)[5])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,6], main = colnames(Brda09.som$codes)[6])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,7], main = colnames(Brda09.som$codes)[7])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,8], main = colnames(Brda09.som$codes)[8])
plot(Brda09.som, type = "property", property = Brda09.som$codes[,9], main = colnames(Brda09.som$codes)[9])

#5. exporting neurons data and codebook vectros data
neurons <- Brda09.som$unit.classif
write.dbf (neurons, "unit_classif.dbf", factor2char = FALSE, max_nchar=254)
codebook_v <- Brda09.som$codes
write.dbf (codebook_v, "codes.dbf", factor2char = FALSE, max_nchar=254)
```

4. SOM mapping script:

```
#1 loading packages and data:
library("kohonen")
library (foreign)
library (plyr)
DATA_ok_07 = read.dbf ("./DATA_07.dbf")
DATA_ok_08 = read.dbf ("./DATA_08.dbf")
DATA_ok_09 = read.dbf ("./DATA_09.dbf")
DATA_ok_10 = read.dbf ("./DATA_10.dbf")
DATA_ok_11 = read.dbf ("./DATA_11.dbf")

#2 selecting variables and scaling data
DATA_ok_07 <- DATA_ok_07 [,c("NET.AREA","Y_PLANTING","B_SPACING","V_SPACING","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM"
,"PARC.AREA")]
DATA_oks_07 <- scale (DATA_ok_un07)
DATA_ok_08 <- DATA_ok_08 [,c("NET.AREA","Y_PLANTING","R_SPACING","V_SPACING","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM"
,"PARC.AREA")]
DATA_oks_08 <- scale (DATA_ok_un08)
DATA_ok_09 <- DATA_ok_09 [,c("NET.AREA","Y_PLANTING","R_SPACING","V_SPACING","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM"
,"PARC.AREA")]
DATA_oks_09 <- scale (DATA_ok_un09)
DATA_ok_10 <- DATA_ok_10 [,c("NET.AREA","Y_PLANTING","R_SPACING","V_SPACING","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM"
,"PARC.AREA")]
DATA_oks_10 <- scale (DATA_ok_un10)
DATA_ok_11 <- DATA_ok_11 [,c("NET.AREA","Y_PLANTING","R_SPACING","V_SPACING","Z_AVG","SLOPE_AVG","EXP_AVG","F.AREA.SUM"
,"PARC.AREA")]
DATA_oks_11 <- scale (DATA_ok_un11)

#3 training SOM from 2009 data:
set.seed (7)
Xtraining <- DATA_oks_09
somnet_09 <- som(DATA_oks_09,
rlen = 60000,
alpha = c(0.05, 0.01),
grid = somgrid(10, 16, "hexagonal"))

#4 plotting data of 2007-2011 on trained SOM from 2009 data:
set.seed(7)
mapping_07 <- map(somnet_09,
scale (DATA_ok_07,
center=attr(DATA_oks_09, "scaled:center"),
scale=attr(DATA_oks_09, "scaled:scale")))
```

```
set.seed(7)
mapping_08 <- map(somnet_09,
scale (DATA_ok_08,
center=attr(DATA_oks_09, "scaled:center"),
scale=attr(DATA_oks_09, "scaled:scale")))

set.seed(7)
mapping_10 <- map(somnet_09,
scale (DATA_ok_10,
center=attr(DATA_oks_09, "scaled:center"),
scale=attr(DATA_oks_09, "scaled:scale")))

set.seed(7)
mapping_11 <- map(somnet_09,
scale (DATA_ok_11,
center=attr(DATA_oks_09, "scaled:center"),
scale=attr(DATA_oks_09, "scaled:scale")))

#5 exporting neurons data for 2007-2011:
somnet_09 <- somnet_09$unit.classif
write.dbf (somnet_09, "somnet_09.dbf", factor2char = FALSE, max_nchar=254)
mapping_07 <- mapping_07$unit.classif
write.dbf (mapping_07, "mapping_07.dbf", factor2char = FALSE, max_nchar=254)
mapping_08 <- mapping_08$unit.classif
write.dbf (mapping_08, "mapping_08.dbf", factor2char = FALSE, max_nchar=254)
mapping_10 <- mapping_10$unit.classif
write.dbf (mapping_10, "mapping_10.dbf", factor2char = FALSE, max_nchar=254)
mapping_11 <- mapping_11$unit.classif
write.dbf (mapping_11, "mapping_11.dbf", factor2char = FALSE, max_nchar=254)
```