Colloquium Biometricum 40
2010, 41–51

EVALUATION OF METHODS FOR THE DETECTION OF SPATIAL OUTLIERS IN THE YIELD DATA OF WINTER WHEAT

Dariusz Gozdowski[1], Stanisław Samborski[2], Eike Stefan Dobers[3]

[1]Department of Experimental Design and Bioinformatics,
Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warsaw, Poland
[2]Department of Agronomy, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776 Warsaw, Poland
[3]Faculty of Geoscience and Geography, Georg-August University,
Goldschmidtstrasse 3, 37077 Göttingen, Germany
e-mails: dariusz_gozdowski@sggw.pl;
stanislaw_samborski@sggw.pl; edobers@gmx.de

Research supported by scientific project MNiSW N N310 089036

Summary

This work presents evaluation of three methods of spatial outlier detection in yield data. Raw yield data used for the analyses came from a field cropped with winter wheat in 2009 located in north of Poland. Three methods were used for the spatial outliers detection, one method based on histogram and two methods based on spatial autocorrelation coefficient (*Moran's I*). Different percentages of the outliers were detected using each of the methods and quite weak correspondence between the methods was achieved.

Key words and phrases: spatial outliers, yield data, winter wheat, spatial autocorrelation

Classification AMS 2010: 62H11

1. Introduction

Farmers use yield maps to target existing resources in areas of low yield to maximise both yield and gross margin. However, the potential to gain a greater financial return and associated possible environmental benefits exists by varying crop inputs to match the yield potential of different parts of the field (Moore and Kremmer, 1998). Yield maps are a valuable source of spatial data in precision agriculture only if they report crop yields close to the actual yields (Faber, 1998).

Unfortunately, devices used to monitor crop yields quite often register data significantly different from actual yield values. Mostly this is due to the dynamics of the movement of grain in the different devices of combine harvester, e.g. change of speed and direction of the harvester as well as improper calibration of yield meters (Colvin and Arslan, 1999, 2002; Arslan and Colvin, 2002). However, the process of harvesting itself very often leads to erroneous data because of logging data without any crop flow in the harvester or harvesting without full header usage.

Despite the fact that yield monitoring technology develops, it is not possible, in the present state, to avoid erroneous yield data registered in some areas of the field. The number of such incorrect data (spatial outliers) saved depends on the presence of obstacles in the field, stops of harvester, etc. Share of the spatial outliers usually ranges from 10 to even 50% (Sudduth and Drummond, 2007).

It is difficult and laborious, to point out the outliers based on raw yield data and visual assessment of yield maps. Statistical methods that could help to detect such outliers are very desirable. The simplest approach, applied to detect erroneous yield data, is the removal of yield values beyond the range of biological potential (Simbahan et al. 2004); in case of cereals yields under Polish conditions this value would be about 12 $t \cdot ha^{-1}$ (COBORU 2008).

To remove outliers it is also possible to apply classical statistical approaches such as removal of yield observation larger or lower by more than 3 standard deviation from the average yield (Simbahan et al. 2004). Such methods of outliers detection allow to remove most of the observation significantly different from the actual yield. However, still a large number of possibly erroneous observations can not be removed. This is because a typical observation in the entire set of values can be a spatial outlier e.g. moderately low value of combine logging data in field areas where yields are very high in reality. This observation is a non-typical observation in space (called spatial outlier), ie. local outlier (Dobermann, 2003; Ping and Dobermann, 2005). Detection of spatial outliers should be based on the method which takes into account the location of the point in space, and assess the values of the yield in the neighboring points. This is possible to achieve if we use spatial statistics (geostatistics) methods.

In case of biological phenomena, (e.g. yield of agricultural crops), very often positive spatial autocorrelation is observed, i.e. the value of a trait is usually similar in the neighboring areas, and changes with increasing the distance (Sokal et al. 1998). The occurrence of spatial autocorrelation, can be checked for example basing on the global spatial correlation, e.g. *Geary* or *Moran's* coefficients. Positive spatial autocorrelation means that the points next to each other usually have similar values. Such positive autocorrelation is almost always observed in case of yield because yields are usually more similar for neighboring points than for distant points (Long, 1998; Robinson and Metternicht, 2005).

Moran's local autocorrelation coefficient relates to the autocorrelation of an individual point. In case where the value of a given point is similar to the values of neighboring points, local autocorrelation coefficient is above 0. If the value of the local autocorrelation coefficient is below 0, this means that a given point value differs much from the neighboring points values. Most methods used for the detection of spatial outliers is based on the existence of such a negative autocorrelation of a single point (McGrath and Zhang, 2003; Shekhar et al. 2003, Anselin et al. 2006). These methods are used for the detection of spatial outliers in various research areas e.g. meteorology, demography, health care, geology, environmental protection and other fields where we are dealing with spatial variables.

To make the detection of spatial outliers easier it seems important to compare the correspondence of results obtained by the method based on local autocorrelation coefficient and classical methods based on non-spatial approach and visual assessment of the map. The latter are probably more objective but require more experience to be done properly.

The aim of this study is to evaluate the usefulness of a method based on local *Moran's I* for the detection of spatial outliers in the yield maps.

2. Material and methods

Raw yield data used for the analyses came from a field cropped with winter wheat (*Triticum aestivum L.*, cv. *Trend*) (21.9 ha) farmed by Farm Frites Poland Dwa Sp. z o.o, located in Bobrowniki (54°52'80"N, 17°32'86" E), Pomerania region, in the north of Poland. The field under study is dominated by brown soils (WRB: Dystric Cambisol), which developed on strong loamy sands from moraine glacial depositions of the last glaciation. According to the Polish system of agricultural suitability categories, these soils are of medium to good suitability for wheat production.

During the harvest in 2009 grain yield was measured by yield monitors mounted on two Claas Lexion 560 harvesters. Raw yield data (3502 points in total) were logged and georeferenced automatically using GPS receiver. A map of raw yield data is presented on Figure 1.

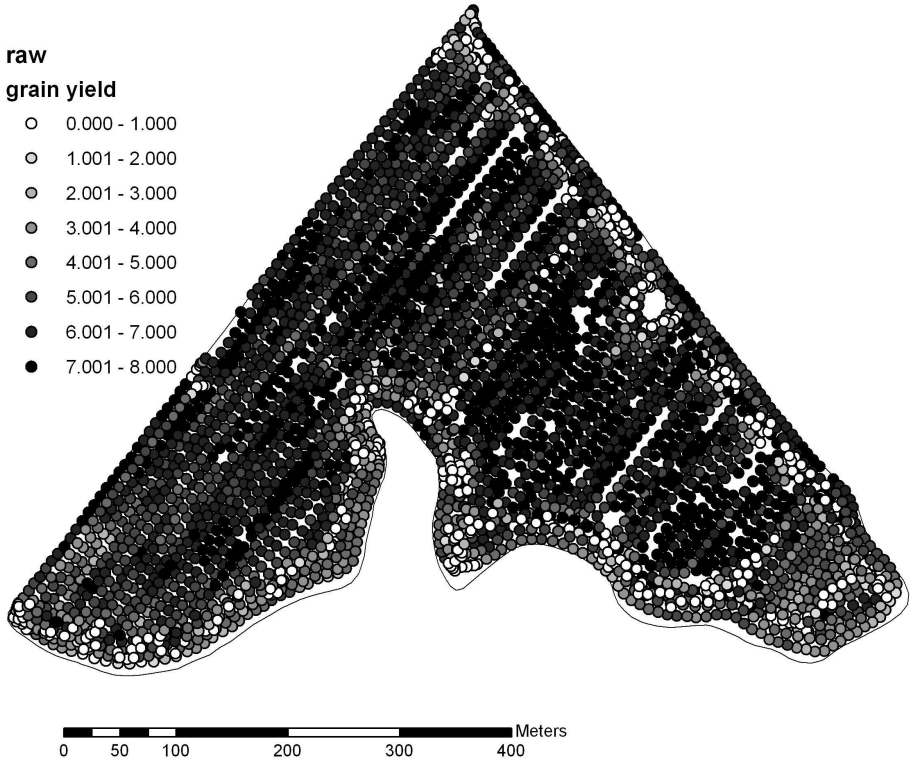

Fig. 1. Map of a raw grain yield of winter wheat

For each point of raw yield data local *Moran's* spatial autocorrelation coefficient was calculated using ArcGIS 9.3 software according to the following formula (Anselin 1995):

$$I_i = \frac{(x_i - \bar{x})\sum_{i=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{N}} \tag{2.1}$$

where: N – number of points, x_i – value of a variable for i-th location, x_j – value of a variable for j-th location, $\overline{x}$ – average value of a variable, w_{ij} – weight between locations i and j. These weights for pairs of points are reciprocals of Euclidean distances between these points. To make possible calculation of local *Moran's I* for each point only points in the radius of 100m were taken into consideration. Most of values *Moran's I* are between -1 and 1 but the values can exceed interval [-1; 1]. Values below 0 mean negative autocorrelation and probability that the point is a spatial outlier.

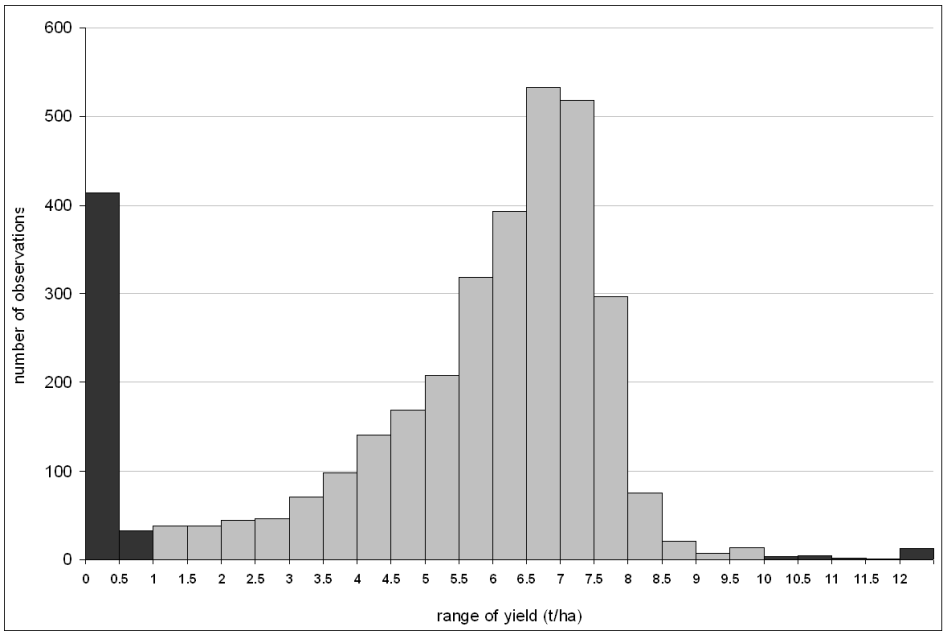

Fig. 2. Histogram for raw yield data (outliers are marked in black)

The same dataset of raw yield data was analyzed using classical (non-spatial) statistics and methods for detection of unusual yield values. For doing so, empirical frequency distributions were calculated for the raw data of the variable *yield* and *harvest rate*, which were available from the logged data file. Thresholds were established on the base of the frequency curve and all data points excluded, which were beyond these thresholds of very low or very high yield and harvest rate, respectively. Figure 2 shows the yield data histogram. Values treated as outliers are mainly in the range from 0 to 1 t ha^{-1} and larger

than 10 t ha^{-1}. However, some data points with yield values within the specified range were marked as outliers because of their non-typical values for the harvest rate, another parameters registered automatically by the yield monitor. The method is straight forward and does not take into account the spatial arrangements of yield data. In this paper, we will refer to this method as histogram method, which led to the exclusion of 692 raw yield data points (19.9%) for further interpolation of yield maps.

3. Results

Yield values were positively correlated, value of global *Moran's I* coefficient was equal to 0.704. It means that positive spatial autocorrelation exists and local Moran's coefficient can be used for outlier detection. For the whole raw yield data set, 758 values (21.6%) with local *Moran's* spatial autocorrelation coefficient below 0 were observed. It means that the value of grain yield for these points is quite different from grain yield observed for the neighboring points because of their negative autocorrelation. In this paper testing of significance of outlier detection was omitted because only 91 points were significant negative autocorrelation (testing based on normal distribution assumption of *Moran's I* coefficient at 0.05 probability level) and we decided to treat all points with value of *Moran's I* below 0 as outliers.

Unfortunately this method for outlier detection only to some extent corresponded with the histogram method because only 103 yield data points were the same outliers for both methods (Fig. 3 and Tab. 1). This result was mainly caused by positive autocorrelation for points located at the edges of the field, where very often low values of grain yield (below 0.5 t per ha) were observed. It is interesting that for points which have grain yield equal to 0 the value of *Moran's I* was frequently high since grain yield for neighboring points was also equal to 0. Therefore we suppose that the method founded on local *Moran's* spatial autocorrelation coefficient alone is not a sufficient and objective method for outlier detection.

To make the detection of spatial outliers, within raw yield data, more objective we proposed a modified method, based not only on negative value of local *Moran's I*, but additionally on very high value of this coefficient. Since the value of *Moran's I* can be beyond the range of [-1; 1] we assumed that *Moran's I* values greater than 4.7 represent not typical points because the value of grain yield is almost the same for theirs neighbors (e.g. is equal 0). The threshold value of 4.7 was estimated on the basis of logistic regression (probit model) where the decision based on the histogram was treated as dependent

binomial variable (Y) (0 – typical value; 1 – outlier) and independent variable (X) values of *Moran's I* (only values greater than 0 were included into dataset for this analysis). Estimated equation of the regression function was as follow:

$$Y=\exp(-3.42+(0.728)X)/(1+\exp(-3.42+(0.728)X)).$$

On the basis of the regression function, *X* value was estimated for which *Y* value is equal to 0.5. We received a value equal to 4.7 and it means that if the value of *Moran's I* was greater than 4.7, the probability that a point is an outlier was greater than 0.5. Using this approach 1137 (32.5% of total) yield data points were treated as a spatial outliers (Tab. 1 and Fig. 4). Moreover, as many as 434 the same yield data points were considered as spatial outliers by the modified and histogram method. With the modified method the detection of spatial outliers was much better than the detection on the basis of negative autocorrelation coefficient alone, but it still gave low correspondence with histogram method.

Table 1. Contingency table presenting number of outliers and typical points for histogram method of outlier detection and two methods based on spatial autocorrelation coefficient – local *Moran's I*

		Method based on negative value of *Moran's I*	
		typical values	outliers
Histogram method	typical values	2155	655
	outliers	589	103
		Method based on negative value of *Moran's I* and very strong autocorrelation (*I*>4.7)	
		typical values	outliers
Histogram method	typical values	2108	702
	outliers	258	434

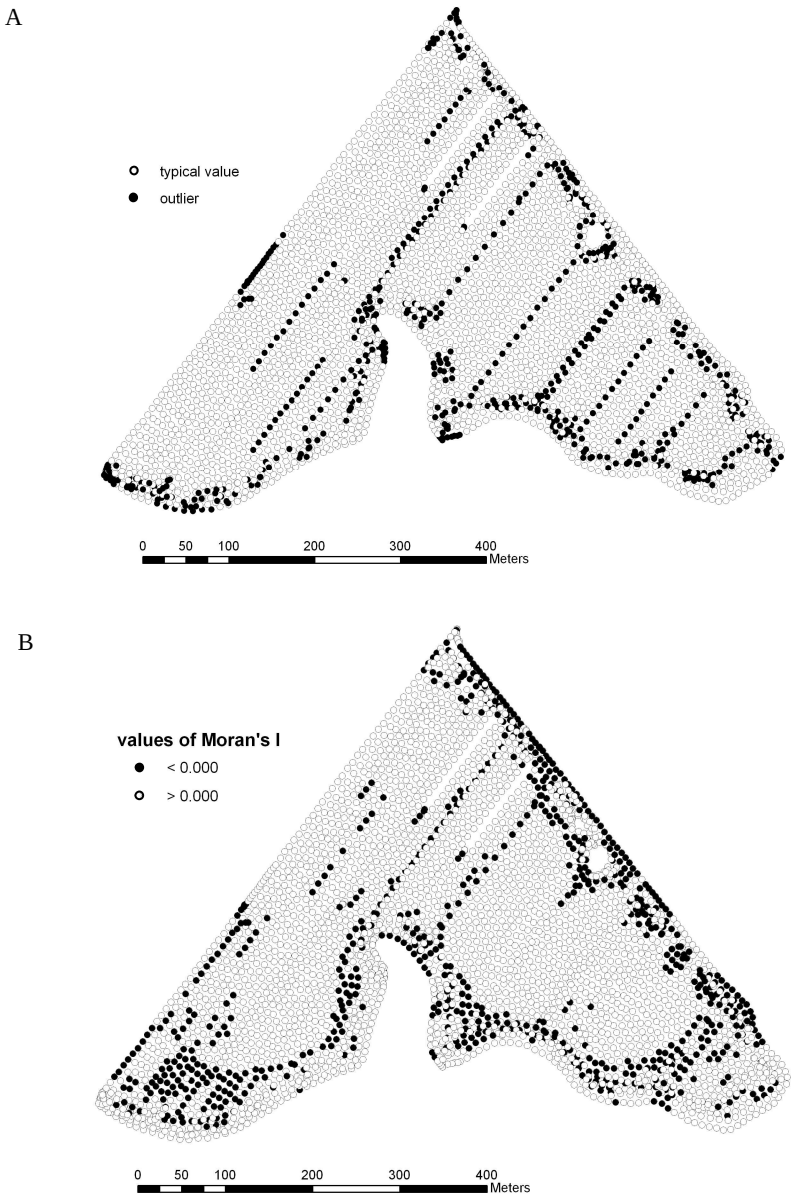

Fig. 3. Outliers detected by histogram method (A) and on the basis of local *Moran's I* (B) (value of *I* below 0 indicates spatial outliers)

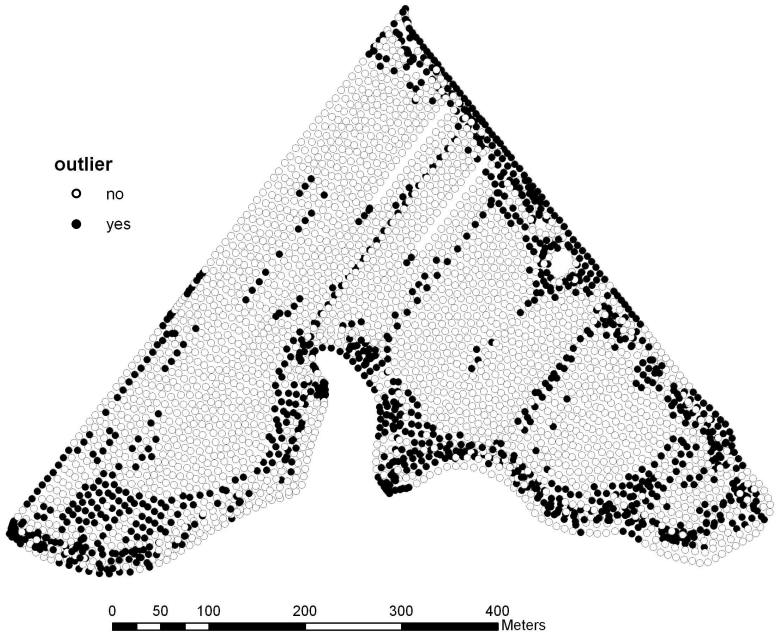

Fig. 4. Spatial outliers detected on the basis of negative value of local *Moran's I* and very high values of *I* (greater than 3)

4. Conclusions

It seems that method of outlier detection based on values of *Moran's I* can be useful especially for detection of individual outliers which are next to typical values. Usefulness of this method is not sufficient if spatial outliers are in groups situated next to each other. Such situation can exist especially in the border of the fields.

This study proved that the use of the autocorrelation coefficient *Moran's I* alone, is not an objective method for the spatial detection of outliers within raw yield data. The detection of spatial outliers based on negative value of *Moran's I* was not sufficient and many outliers pointed out earlier by the histogram method were not detected.

It has been observed that not only negative autocorrelation coefficient *Moran's I* but also its very high value can be the indicator of an outlier.

The process of detection of spatial outliers should consist of classical methods (e.g. removing very high and very low values of grain yield) and complementary methods based on the autocorrelation coefficient as a final step for creation of reliable yield maps.

References

Anselin L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis* 27, 93–115.

Anselin L., Syabri I., Kho Y. (2006). GeoDa: An introduction to spatial data analysis, *Geographical Analysis* 38, 5–22.

Arslan S., Colvin T. (2002). Grain yield mapping: yield sensing, yield reconstruction, and errors. *Precision Agriculture* 3, 135–154.

COBORU (2008). Wyniki Porejestrowych Doświadczeń Odmianowych w województwie pomorskim, SDOO Radostowo.

Colvin T., Arslan S. (1999). *Yield monitor accuracy*. SSMG-9, Potash & Phosphate Institute, 1–4.

Colvin T., Arslan S. (2002). An evaluation of the response of yield monitors and combines to varying yields. *Precision Agriculture* 3, 107–122.

Dobermann, A., Ping J.L., Adamchuk V.I., Simbahan G.C., Ferguson R.B. (2003). Classification of crop yield variability in irrigated production fields. *Agronomy Journal* 95, 1105–1120.

Faber A. (1998). System rolnictwa precyzyjnego. I. Mapy plonów. *Fragmenta Agronomica* 1(57), 4–15.

Long D.S. (1998). Spatial autoregression modeling of site-specific wheat yield. *Geoderma* 85, (2–3), 181–197.

McGrath D., Zhang C. (2003). Spatial distribution of soil organic carbon concentrations in grassland of Ireland. *Applied Geochemistry* 18,1629–1639.

Ping J. L., Dobermann A. (2005). Processing of yield map data. *Precision Agriculture* 6, 193–212.

Moore M.R., Kremmer P.R. (1998). An investigation into the factors influencing the accuracy of yield maps. The International Fertiliser Society - Proceeding 421.

Robinson T.P., Metternicht G. (2005). Comparing the performance of techniques to improve the quality of yield maps. *Agricultural Systems* 85 (1), 19–41.

Shekhar S., Lu C.T., Zhang P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica* 7 (2), 139–166.

Simbahan G.C., Dobermann,A., Ping J.L. (2004). Screening yield monitor data improves grain yield maps. *Agronomy Journal* 96, 1091–1102.

Sokal R.R., Oden N.L., Thomson B.A. (1998). Local spatial autocorrelation in biological variables. *Biological Jounal of the Linnean Society* 65, 41–62.

Sudduth K., Drummond S. (2007). Yield Editor: Software for removing errors from crop yield maps. *Agronomy Journal* 99, 1471–1482.

OCENA METOD DETEKCJI OBSERWACJI ODSTAJĄCYCH W PRZESTRZENI W DANYCH DLA PLONÓW PSZENICY OZIMEJ

Streszczenie

W pracy przedstawiono ocenę trzech metod detekcji obserwacji odstających w plonach pszenicy ozimej. Dane do analiz pochodziły z pola, na którym uprawiano pszenicę w roku 2009 położonym na północy Polski. Zostały wykorzystane trzy metody detekcji obserwacji odstających (jedna metoda oparta na histogramie oraz dwie metody oparte na współczynniku autokorelacji przestrzennej (*I Morana*). Uzyskano różny udział procentowy obserwacji odstających oraz dość niewielką zgodność wyników uzyskanych ocenianymi metodami.

Słowa kluczowe: obserwacje odstające w przestrzeni, plony roślin, pszenica ozima, autokorelacja przestrzenna

Klasyfikacja AMS 2010: 62H11