Paweł Kobus[1]
Chair of Agricultural Economics and International Economic Relations
Warsaw University of Life Sciences
Warsaw, Poland

Modelling wheat yields variability in Polish voivodeships

Abstract. The paper presents an analysis of wheat yields variability in the voivodeships of Poland. The main aim of the study was to find out what are the statistical relationships between the wheat yield variability and the following factors: arable area, size of wheat production area, share of arable land used for wheat production, land quality and average yield. For that purpose a multiple linear regression was applied. It was found out that the detected spatial autocorrelation of wheat yields variability measured by standard deviations can be explained in 75% by the fitted model. Two of the considered variables showed a significant negative effect on this variability: the logarithm of arable area and the land quality, while the other two: the average wheat yield and the wheat production area displayed a significant positive effect on the variability. The effect of share of arable land used for wheat production itself was not significant.

Key words: wheat yield, variability, production risk, spatial autocorrelation.

Introduction

The production risk is one of the dominant risks in the agricultural sector, particularly in plant cultivation. The most important factors influencing yield level are weather conditions, pest and diseases. Poland is a country of a relatively moderate size but the weather patterns and soil conditions are quite diverse. The same can be said about the interaction of weather and soil conditions and for that reason an aggregation should be avoided when assessing production risk in plant cultivation. Unfortunately voivodeships are the smallest area entities for which yield time series of reasonable length are available.

In previous work by the author [Kobus 2009] it was shown that the standard deviation for wheat yield (after trend elimination) takes values from 2 decitons per hectare (dt/ha) in Podkarpackie voivodeship to 5.4 dt/ha in Lubuskie voivodeship. It proves that even on the NUTS 2 level there exists a considerable unevenness of plant production risk, which is worth of analysis.

Such an irregularity of yield variability may be a result of many factors. Grønlund et al. [2006] found that 19% of wheat yield variability in Norway (at farm level) can be explained by 15 significant predictors, the highest ranking predictors were irrigation, winter wheat percentage in wheat sown area, pH and natural logarithm of farm area.

Other authors [Górski & Górska 2006] investigated a relationship between the level of data aggregation, from field plot data to national yields, and the yield variability. They found it negative.

The reason for the relationship between production area and yield variability can be explained on the ground of probability. Assuming that yield from each hectare is a random variable, an average of such variables is also a random variable but with a variance equal to

[1] DrSc, email: pawel_kobus@sggw.pl.

the average variance of individual variables only in the case of full linear dependency between individual yields. As such perfect linear dependency is rarely observed in practice, the variance of an average yield on bigger number of hectares should be smaller.

On the other hand, increasing the production area moves the marginal field to less favourable area and it could result not only in lower yields but also in their higher variability. This could conceal or even negate the effect of larger production area. Consequently, to find out what are the effects of production area size and of incorporating less favourable area, both effects must be studied jointly.

Another factor which the above mentioned unevenness of plant production risk can be attributed to is the average yield level. It has been shown [Kobus 2010] that an average yield level displayed a positive relation with the wheat yield variability at country level. But in the subject matter literature, contradictive results can be found. Haberle and Mikysková [2006] carried out research in the Czech Republic at district level. They investigated 77 administrative units and found out fairly strong (r=-0.58) negative relation between the average wheat yield and its variability measured by the variation coefficient.

The main aim of this study is to find out what are the relationships between wheat yield variability and the following factors: size of arable land, size of wheat production area, share of arable land used for wheat production in the total arable land, land quality and average yield.

Apart from that a possibility of spatial correlation between yields variability in particular voivodeships will be investigated.

Data and research methods

The statistical data used in this analysis include the average yields of wheat in Polish voivodeships in years 1995-2007 and are available from Eurostat [Eurostat 2010], another source of data were the Central Statistical Office of Poland (GUS) [Rocznik… 2010] and Global Administrative Areas [GADM… 2010].

The following variables were used in the analysis:

AA – arable area, ‘000 ha

WPA – wheat production area, ‘000 ha

SPA – share of wheat production area in the arable land area, %

LQ – land quality

AY – average yield of wheat, dt/ha

SD – standard deviation of wheat yield after the linear trend elimination, dt/ha.

Apart from the arable area and wheat production area which were directly available from the Eurostat, all other variables were calculated by author basing on the Eurostat and GUS data. The description given above is sufficient for almost all variables but LQ. This is an artificial variable created for describing the agricultural quality of land. It is a compromise between the available data and accuracy.

In the statistical yearbook by GUS [Rocznik… 2010] on page 75, a table entitled ‘Agricultural land by soil valuation classes and by voivodeships in 2000’ is presented, with agricultural land classified by quality classes, with subclasses a and b in each class pooled. The LQ variable was calculated on the base of that table and coefficients for recalculation of actual into standardised quality hectares employed for taxation purposes in macroregion 2 (class I coefficient 1.8, II – 1.65, IIIa – 1.5, IIIb – 1.25, IVa – 1, IVb – 0.75, V – 0.3, VI –

0.15), while average coefficients for subclasses a and b were used as coefficients for land classes III and IV.

Table 1. Values of analysed data in Poland's voivodeships (1995-2007)

Voivodeship	AA, '000 ha	WPA, '000 ha	SPA, %	LQ	AY, dt/ha	SD, dt/ha
Dolnośląskie	864.8	283.9	32.8	0.933	40.6	4.5
Kujawsko-Pomorskie	972.6	197.3	20.3	0.868	38.4	3.7
Lubelskie	1275.3	296.0	23.2	0.923	31.4	2.5
Lubuskie	392.3	56.9	14.5	0.674	32.9	5.4
Łódzkie	958.9	103.2	10.8	0.664	30.7	3.4
Małopolskie	553.4	118.3	21.4	0.847	30.1	2.7
Mazowieckie	1641.9	156.0	9.5	0.665	31.2	2.6
Opolskie	491.8	154.3	31.4	0.939	45.9	4.7
Podkarpackie	587.9	130.6	22.2	0.845	30.0	2.0
Podlaskie	759.5	60.5	8.0	0.605	27.4	3.3
Pomorskie	689.0	142.5	20.7	0.786	41.6	3.6
Śląskie	415.2	67.4	16.2	0.740	34.8	4.1
Świętokrzyskie	515.4	89.1	17.3	0.820	28.6	3.2
Warmińsko-Mazurskie	835.9	153.7	18.4	0.811	36.8	2.3
Wielkopolskie	1542.5	218.9	14.2	0.702	40.4	4.7
Zachodniopomorskie	856.0	191.2	22.3	0.799	37.8	4.3

Source: own calculations based on [Eurostat… 2010] and [Rocznik… 2010].

For the analysis of data, two models were considered, a standard linear model of multiple regression and one of the models specific to spatial data analysis, namely simultaneous autoregressive model [Bivand et al. 2008]. The model of first choice was a linear model given by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where: $\mathbf{Y}$ – vector of dependent variables (standard deviations of yield SD), $\mathbf{X}$ – matrix of independent variables[2] (log(AA), WPA, SPA, LQ, AY), $\boldsymbol{\beta}$ – vector of regression coefficients, $\boldsymbol{\varepsilon}$ – vector of independent identically normally distributed random errors.

The problem which arises in the application of the above model to the analysed data is the independency of random errors. As the observations are area entities, it is possible that they are spatially autocorrelated. In such case it is inappropriate to use model (1) which assumes independency of random errors.

One of the most popular tests for detecting spatial autocorrelation is the Moran I test [Bivand et al. 2008]:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} * \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{2}$$

[2] The function log(x) is the natural logarithm of x.

where w_{ij} is the spatial weight of the link between areas i and j.

The matrix of spatial weights was constructed on the base of object's neighbourhood presented in Figure 1[3].

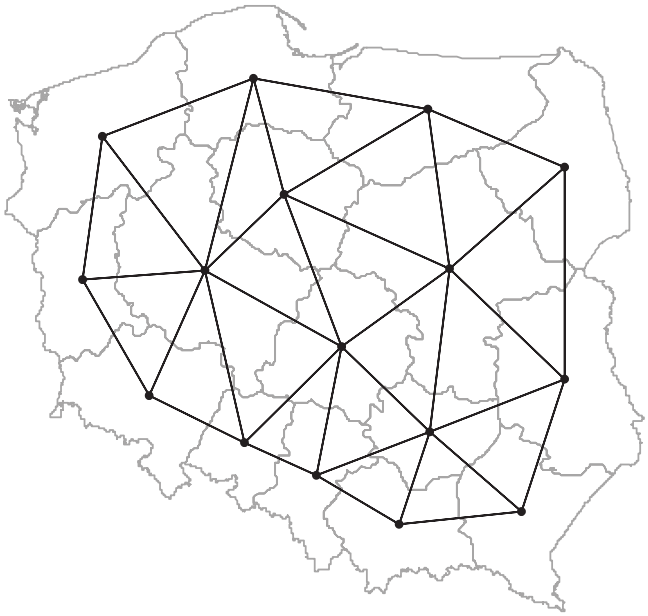

Fig. 1. Neighbourhood structure of Polish voivodeships

Source: own calculations, based on [GADM… 2010].

All of the statistical calculations were performed in R, an environment for statistical computing [R. A language… 2009] with help of the following packages: sp, rgdal, maptools, RcolorBrewer, classInt, spdep.

Results

The spatial distribution of standard deviation of wheat yield after the linear trend elimination, as shown in Figure 2, suggests that there exists a spatial autocorrelation of yield variability. It is clear that all the western voivodeships exhibit comparatively high values of wheat yield standard deviation, while the eastern voivodeships demonstrate low variability. This kind of clustering is typical for objects spatially correlated.

To confirm this visual impression and to quantify this spatial autocorrelation the Moran I test was applied. The value of Moran I statistic was 0.544 with p-value 0.0000113, which proves that variability of wheat yields expressed by their standard deviations is positively spatially correlated, i.e. a voivodeship with high yield variability is likely to have neighbours with similarly high variability.

Although the existence of spatial autocorrelation of yield variability was confirmed, it does not mean that a standard linear model can not be applied. The assumption in the model

[3] The weight matrix in binary style, units for neighbouring voivodeships, otherwise zeros.

concerned the random errors and not the dependent variable itself. It is possible that a large part or even the whole of spatial autocorrelation can attributed to analysed factors.

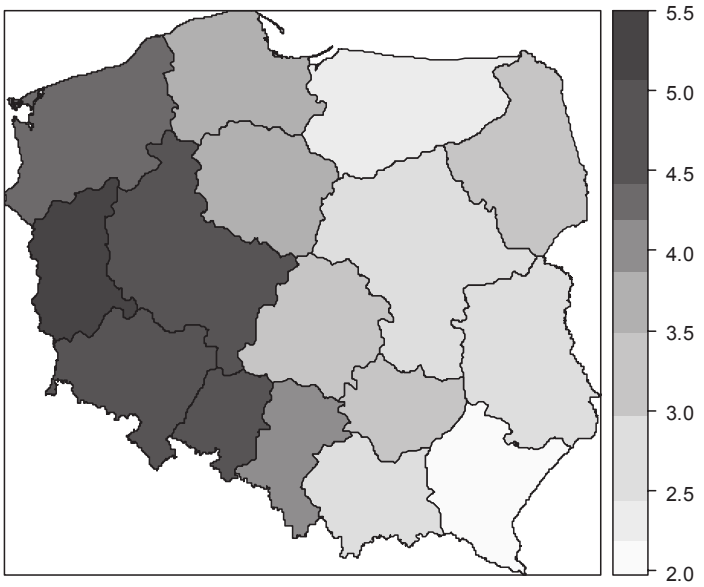

Fig. 2. Standard deviation of wheat yield void of trend, dt/ha (1995-2007).

Source: own calculations.

The results of multiple linear[4] regression model estimation are presented in Table 2. As presented, all variables but the share of wheat production area (SPA) display a significant relation to yield variability at significance level 0.05 and the land quality (LQ) even at that of 0.01.

Table 2. Results of multiple linear regression estimation

Variable	Coefficient estimate	Standard error	t value	p-value
(Intercept)	28.055	8.692	3.228	0.0091
log(AA)	-3.223	1.183	-2.723	0.0214
WPA	0.020	0.007	2.755	0.0203
SPA	0.207	0.663	0.312	0.7613
LQ	-12.920	3.534	-3.656	0.0044
AY	0.114	0.042	2.702	0.0222

Multiple R-squared: 0.7482, adjusted R-squared: 0.6222

Source: own calculations

The results of the used tests are substantial only if the model assumptions hold. To verify the assumption of random errors independence, the Moran I test for linear models

[4] The name linear model does not imply that conditional expected value of dependent variable is a linear function of independent variable. It means that is a linear function of model parameters.

was applied. The value of Moran I statistic was -0.0867 with p-value 0.2231 for hypothesis of null spatial autocorrelation. It means that the spatial dependence of yield standard deviation detected in the previous test can be explained by the investigated factors.

Considering the value of multiple R-squared equal to 0.7482, the estimated regression function explains a large part, almost 75%, of variability of wheat yield standard deviations across Polish voivodeships. For testing of the second assumption (normal distribution of random errors) the Shapiro test was used, with the p-value for hypothesis of normality equal to 0.3438. Concluding, all assumptions of model (1) were confirmed.

Because the share of wheat production area (SPA) was not significant, the variable SPA was removed and the model was re-estimated. Results are presented in Table 3.

Table 3. Results from multiple linear regression without the variable SPA.

Variable	Coefficient estimate	Standard error	t value	p-value
(Intercept)	26.662	7.147	3.731	0.0033
log(AA)	-2.970	0.827	-3.593	0.0042
WPA	0.019	0.006	2.951	0.0132
LQ	-12.601	3.241	-3.888	0.0025
AY	0.106	0.033	3.240	0.0079

Multiple R-squared: 0.7457, adjusted R-squared: 0.6532

Source: own calculations.

Removing the variable SPA resulted in lowering multiple R-squared by only 0.0025 but in increasing adjusted R-squared from 0.6222 to 0.6532. All others variables remained significant with even lower p-values.

All considered factors show a fairly similar strength of influence on the yield standard deviation, but they can be ordered by absolute value of t statistics starting from the strongest: land quality (LQ), natural logarithm of arable area (log(AA)), average yield (AY) and wheat production area (WPA).

The most important are the signs of estimates, they inform on character of relation. A minus sign denotes a negative relationship while a plus sign a positive one. But before jumping to conclusions like 'the bigger wheat production area the bigger average yield variability' one thing must be made clear: the estimates presented in the Tables 2 and 3 come form a multiple regression model and their proper interpretation is that if, for instance, the average yield for a voivodeship increases by one unit (dt/ha), **and all other variables remain on the same level**, the yield standard deviation for that voivodeship will increase by 0.106 dt/ha. Bearing that in mind the conclusion 'the bigger wheat production area the bigger average yield variability' is an oversimplification, the bigger wheat production area without increasing total arable area means higher share of wheat production area and consequently expanding wheat production to less favourable areas. On the other hand the minus before estimate of the log(AA) coefficient means that increasing total arable area lowers yield variability, but it could also be an effect of a relatively decreasing share of wheat production area.

To asses the combined effect of increasing both arable area and wheat production area, the log transformation of arable area (AA) must be taken into account. An increase of arable area by 1 thousand hectare gives a higher increase of log(AA) for a lower reference

point then for higher. The log transformation simply results in constant reaction of dependent variable to increase of arable area (AA) expressed in percentages, while not transformed wheat production area (WPA) causes a constant reaction of depended variable to an increase of WPA expressed in thousands of hectares.

Discussion

The analysis of yield variability showed its significant relationship to 4 out of 5 considered factors, namely: land quality, natural logarithm of arable area, average yield and wheat production area. The share of wheat production area in total arable land was not significant.

The size of arable area and the land quality showed a negative relationship with the wheat yield variability while the wheat production area and the average wheat yield in a voivodeship showed a positive relationship. Those 4 variables explain together nearly 75% of wheat yield standard deviation unevenness among voivodeships. This is a very high value. In a research carried out in the Czech Republic [Haberle & Mikysková 2006], it was 35% and in Norway only 19%. The reason for such a high value of determination coefficient as observed could be the level of data aggregation. In the present study it was a voivodeship with an average size of arable area of 834.5 thousand hectare, while in the Czech Republic it was a district with an average size of arable area of 37.8 thousand hectare and in Norway it was estimated at the farm level.

The negative influence of arable area size on yield variability showed in this study agrees with results presented by Górski & Górska [2006]. Also, the negative influence of land quality agrees with results by Haberle & Mikysková [2006], where a positive influence of high proportion of fertile soils on an average yield and a negative one on its variability was shown.

In a previous paper by author [Kobus 2010] it was shown that an average yield level displays a positive relation with the wheat yield variability at country level. This study confirms that it is true also at the level of provinces, at least in Poland. In the Czech Republic [Haberle & Mikysková 2006], it was found out that such relationship is also significant but negative.

Those results are not necessarily contradictive. In present study variability was measured by standard deviations while Haberle and Mikysková used variation coefficient. Although variation coefficient is a well established and widely used measure of variability, it should not be used (in author's opinion) when investigating the possible relation of the average level of yield and the yield variability. The variation coefficient is a ratio of standard deviation to the average and consequently variation coefficients tend to be negatively correlated with averages even if standard deviations are not.

Conclusions

The results of the analysis confirm that an increase of wheat production area results in lower yield variability. But to have such a result this increase must be connected with expanding the total arable area. An increase of wheat production area achieved by simply

increasing its share in the arable land area will result in an increase of yield variability, which is supposedly a result of moving production to less favourable areas.

An increase of average yield results in a higher yield variability (in absolute terms). The land quality had a reducing effect on wheat yield variability.

The use of multiple regression allowed to determine an unmasked effect of considered factors and proved to be useful for modelling wheat yields variability in Polish voivodeships. Although in case of other crops or a different level of data aggregation, a possible spatial autocorrelation should be considered.

References

Bivand R.S., Pebesma E.J., Gómez-Rubio V. [2008]: Applied Spatial Data Analysis with R. Springer, New York.

Eurostat [2010]. Statistical Office of the European Communities. [Available at:] http://ec.europa.eu/eurostat. [Accessed: June 2010].

GADM. Database of Global Administrative Areas. [2010]. [Available at:] http://gadm.org/ [Accessed June 2010].

Górski T., Górska K. [2006]: The effects of scale on crop yield variability. *Agric. Systems* 78, pp. 425-434.

Grønlund A., Njøs A., Vestgarden L.S., Lyngstad I. [2006]: Utilisation of agricultural databases for statistical evaluation of yields of barley and wheat in relation to soil variables and management practices. *Acta Agric. Scand.* B 56, pp. 1-8.

Haberle J., Mikysková J. [2006]: Relation of cereals yields and variability to soil-climate and production characteristics of districts of the Czech Republic. *Journal of Central European Agriculture*. vol. 7 no. 4, pp. 661-668.

Kobus P. [2009]: Wheat yields variability in Poland at NUTS 2 level in context of production risk. *Scientific Journal of Warsaw University of Life Sciences* series *Problems of World Agriculture*, vol. 6 (XXI), pp. 51-58.

Kobus P. [2010]: Changes of level and variability of wheat production in EU Member States, period 1961-2008. *Economic Science for Rural Development* no. 21, pp. 90-99.

R: A language and environment for statistical computing. [2009]. R Development Core Team. R Foundation for Statistical Computing, Vienna, Austria. [Available at:] http://www.R-project.org. {Accessed: September 2010].

Rocznik Statystyczny Rolnictwa 2009 [2010]. Zakład Wydawnictw Statystycznych GUS, Warsaw.