**Paweł Kobus**[1]
Department of Agricultural Economics and International Economic Relations
Warsaw University of Life Sciences – SGGW

# Modelling joint distribution of crop plant yields and prices with use of a copula function

**Abstract.** The paper constitutes an attempt at modelling the joint distribution of crop plant yields and prices in Poland. The main objective of the paper was to examine the usefulness of the copula function for the task and the selection of suitable marginal distributions. The fit of a joint distribution based copula function was compared with multivariate normal distribution. It was revealed that the multivariate normal distribution is outperformed by a Gaussian copula with the following marginal distribution: yields of both crop plants – normal distribution, price of wheat – Burr distribution (type XII) and price of rapeseeds – lognormal distribution. The main advantages of the copula function were: the possibility to use different marginal distributions and ability to model non-elliptical two-dimensional distributions. The practical implications of choosing the right joint distribution is demonstrated by comparing empirical quantiles of income for a given crop structure with theoretical quantiles based on the proposed joint distributions.

**Key words:** joint distribution, yields and prices, income risk, copula function

## Introduction

Income risk in agriculture is most strongly affected by crop plant yields and prices. To properly evaluate the income risk of the crop structures examined, one should calculate at least the first two moments of the income generated by this crop structure, that is, a sum of yield-price products. The calculation of income distribution moments must be preceded by an estimation of the joint multi-dimensional distribution of crop plant yields and prices.

It has so far been assumed that the relation between yields and prices of the entire group of the plants being examined is explained sufficiently well enough by a correlation matrix. Consequently, it was believed that the multidimensional distribution of yields and prices can be sufficiently approximated by a multivariate normal distribution.

Regrettably, this strong assumption is not justified even in case of a marginal distribution [Tejeda and Goodwin 2008]. It cannot be expected that each of the examined variables follows normal distribution or even in fact, the same distribution. Therefore, it is reasonable to look for such a tool that will allow to incorporate various marginal distributions into one joint distribution of yields and prices [Zhu et al. 2008, Schulte-Geers and Berg 2011].

This paper aims at verifying the usefulness of a copula function for modelling joint distribution of crop plant yields and prices in Poland and for the selection of suitable marginal distributions.

[1] PhD, e-mail: pawel_kobus@sggw.pl

## Data

This analysis uses farm level data from the Polish Farm Accountancy Data Network (FADN). The process of data selection was as follows: samples from years 2004 – 2009 were screened for farms which were present in the samples in all the years, and for which yields and transaction data for winter wheat and rape were available for all the years examined. In the end, a sample consisting of 378 farms was selected.

Observations of the following variables were available for each farm:

$X_1$ – winter wheat yield [dt/ha];

$X_2$ – rape yield [dt/ha];

$X_3$ –wheat price [PLN/dt];

$X_4$ – rapeseeds price [PLN/dt].

Observations from all the farms and from all years were analysed together. Thus, 2268 repetitions were obtained for each variable.

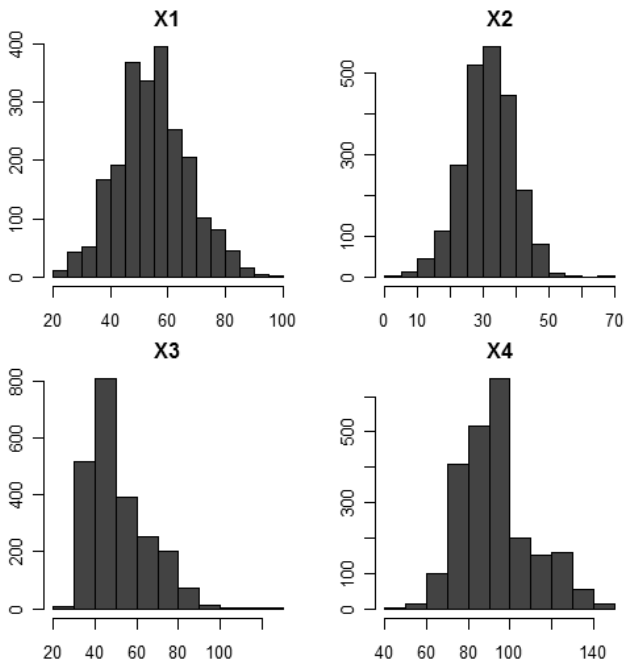


Fig. 1. Marginal distributions of yields and prices for winter wheat and rape

Source: own calculations, based on FADN data

The histograms in the Fig. 1 confirm that the shape of the distribution is relatively close to normal distribution only for yields ($X_1$ and $X_2$). The prices, especially those of wheat ($X_3$), manifest a positive skew which is too high for a normal distribution. The values of descriptive statistics in Table 1 also support the first impression about yield and price distributions. For the yields ($X_1$ and $X_2$), kurtosis is very close to 3 and the skewness coefficient is close to 0, while for wheat prices ($X_3$) skewness is 1.03 and for rapeseed ($X_4$) it is 0.65.

Table 1. Basic characteristics of the yield and price distributions

| Descriptive statistics | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| Average | 55.88 | 31.79 | 51.13 | 92.86 |
| Standard deviation | 12.29 | 7.86 | 14.25 | 16.85 |
| Variation coefficient | 0.220 | 0.247 | 0.279 | 0.182 |
| Median | 55.00 | 32.00 | 47.15 | 90.94 |
| Kurtosis | 2.99 | 3.26 | 3.81 | 3.15 |
| Skewness | 0.15 | -0.18 | 1.03 | 0.65 |

Source: own calculations, based on FADN data

On the basis of the results from Table 1, it was decided to consider 3 marginal distributions: normal, lognormal and Burr (type XII), the last one allows for extreme right skewness and is a good candidate for $X_3$ and $X_4$.

## Methods

We start the process of searching for an appropriate joint distribution of yields and prices by considering options for marginal distributions, than we estimated dependence structure of joint distribution using Gaussian copula function. To compare various distribution Voung test [Voung 1989] was applied.

Density function of normal distribution $N(\mu,\sigma^2)$:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

Density function of lognormal distribution $LN(\mu,\sigma^2)$:

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ x>0. \tag{2}$$

Density function of three-parameter Burr[2] (type XII) distribution $Burr(\alpha,\tau,\varphi)$:

$$f(x)=\tau\alpha\left(\frac{x}{\varphi}\right)^{\tau}\Bigg/\left(x\left(1+\left(\frac{x}{\varphi}\right)^{\tau}\right)^{\alpha+1}\right),\ x>0,\alpha>0,\tau>0,\varphi>0. \tag{3}$$

[2] See [Tadikamalla 1980] for a friendly introduction to Burr distribution.

For modelling the joint distribution copula function was applied, where p-dimensional copula $C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$ is defined as multi-dimensional distribution on $[0, 1]^p$ space, with marginal distributions following standard uniform distribution U(0,1). It was proved in [Sklar 1959] that any multi-dimensional distribution $F(x_1, x_2,\ldots, x_p)$ with marginal distributions functions $F_1, F_2, \ldots, F_p$ can be written as follows:

$$F(x_1, x_2, \ldots, x_p) = C\left(F_1(x_1), F_2(x_2), \ldots, F_p(x_p); \boldsymbol{\theta}\right) \tag{4}$$

where $\boldsymbol{\theta}$ is copula function parameters vector.

In this paper, the multi-dimensional distribution was estimated as follows: first, the marginal distribution was estimated using the maximum likelihood method, then next, for the selected type of copula function, i.e., Gaussian copula, dependency parameters were estimated using the maximum pseudo-likelihood method. In case of Gaussian copula, the parameters vector $\boldsymbol{\theta}$ is a vector of correlations $[\rho_1, \rho_2, \ldots, \rho_k]$, where $k = ½\, p^2 - p$.

When we consider two or more models for describing the distribution of an observed variable, we need a procedure for choosing this model, which is significantly better. One popular approach is to use the likelihood ratio (LR) test. However, the LR test can be used only when the models being compared are nested. Using the Kullback-Leibler information criterion, Voung proposed the closeness likelihood ratio based test for non-nested models [Voung 1989]:

$$z_V = \frac{LL_{\hat{A}} - LL_{\hat{B}} - \frac{p_A - p_B}{2}\log(N)}{\sqrt{N\hat{\omega}^2}} \tag{5}$$

where $LL_{\hat{A}}$ and $LL_{\hat{B}}$ are log-likelihoods of estimated models A and B, $p_A$ and $p_B$ are numbers of their parameters, N is the number of observations and $\hat{\omega}^2$ is sample variance of the pointwise log-likelihood ratios. According to theorem 5.1 in [Voung 1989]:

- under the $H_0$ (the null hypothesis about both models being equally close or distant from the true model), the $z_V$ statistic follows standard normal distribution N(0,1);
- under the $H_A$, that is, the alternative hypothesis that model A is closer to the true model, $z_V \to \infty$;
- and under the $H_B$, that is, the alternative hypothesis that model B is closer to the true model, $z_V \to -\infty$.

This theorem provides a simple rule for deciding which model is better: if $z_V > c$ then model A is significantly better than model B, and if the value of $z_V < -c$ then model B is the better one, where $c$ is a critical value from standard normal distribution of a chosen significance level.

The calculations for all models were performed in R, a statistical computing environment [R Core Team 2013] with help of the 'copula' package [Hofert et al. 2013] and the 'actuar' package [Dutang et al. 2008].

## Results

As already mentioned, in this paper there are 3 distributions: normal, lognormal and Burr (type XII), which are considered as options for marginal distributions. All three were fitted for each of variables: $X_1$, $X_2$, $X_3$ and $X_4$. Next, Voung test was used for selecting the best one in each case.

Table 2. Results of Voung test for the yield and price distributions

| Compared distributions | Values of $z_V$ statistics | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| Burr v. Normal | -1.319 | -1.368 | 8.927 | 6.012 |
| Burr v. Log-normal | 3.757 | 5.252 | 1.756 | -1.742 |
| Normal v. Log-normal | 3.836 | 5.092 | -16.032 | -10.385 |

Source: own calculations, based on FADN data

The interpretation of values in Table 2 need some clarification. For example, in the first line, when comparing Burr and normal distributions, we see 6.012 in the last column, which means that for variable $X_4$, the Burr distribution is closer to the true model than normal distribution. What it is more, the value 6.012 compared with the 95% quantile of the standard normal distribution (1.6448) proves that this is a significant difference. But if we look at the second row where Burr and log-normal distribution are being compared, we see the $z_V$ statistic with the value of -1.742, meaning that the Burr distribution is significantly farther from the true one than the log-normal distribution.

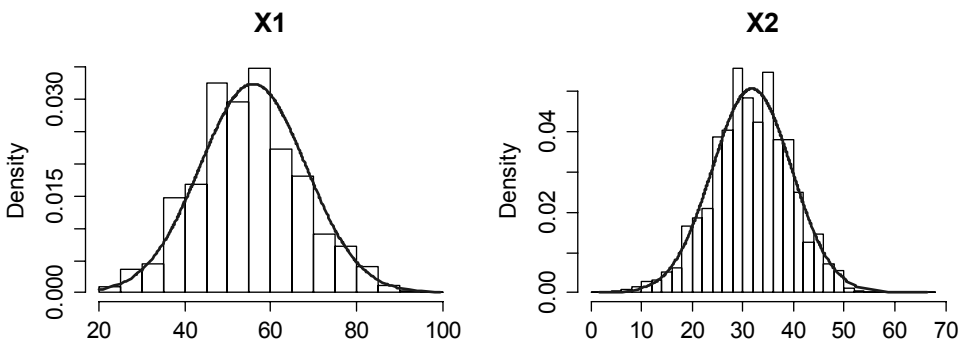


Fig. 2a. Fitted marginal distributions of yields for winter wheat and rape

Source: own calculations, based on FADN data

In the end, following distributions were selected: $X_1 \sim N(55.880, 12.295)$, $X_2 \sim N(31.792, 7.857)$, $X_3 \sim Burr(0.305, 12.530, 39.234)$, $X_4 \sim logN(4.515, 0.178)$, the values given in parentheses being maximum likelihood estimators of distribution parameters.

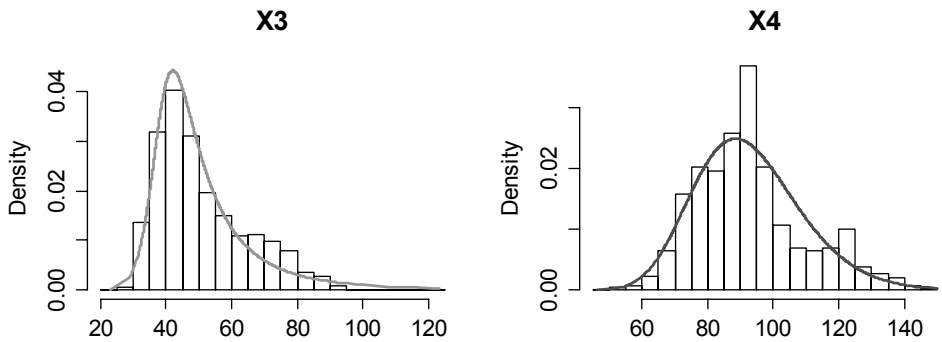


Fig. 2b. Fitted marginal distributions of prices for winter wheat and rape

Source: own calculations, based on FADN data

In Fig. 2a and Fig. 2b we can see, that except for the price of rapeseed ($X_4$), all other density functions seem to fit the empirical data rather well. Nevertheless, these were only marginal distributions. It is not possible to depict on paper a distribution above a dimension of 2. Fig. 3 shows the scatterplots for each combination of variables, which at least makes it possible to see the 2-dimensional relation between variables

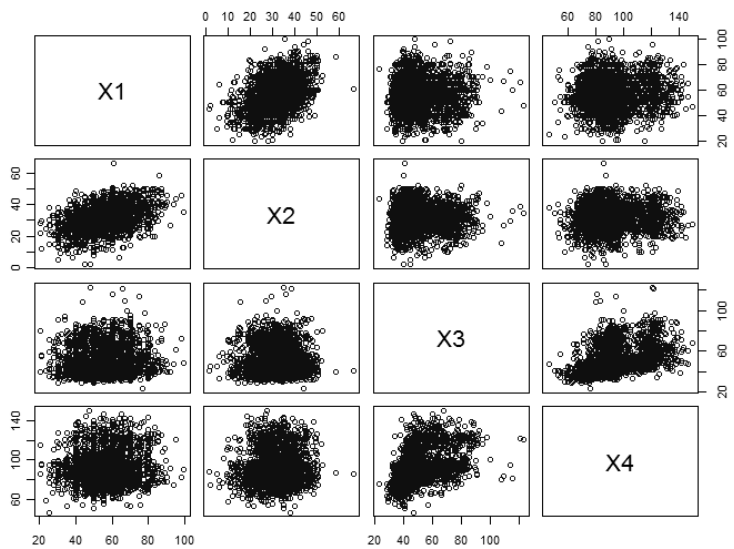


Fig. 3. Two-dimensional scatterplots for joint distribution of yields and prices for winter wheat and rape

Source: own calculations, based on FADN data

It was evident that only scatterplots for the 2-dimensional distribution of $X_1$ and $X_2$ have the typical elliptical shape of a bivariate normal distribution (see graphs in Fig. 3: first row, second column or second row, first column). In the remaining cases, especially for $X_3$ and $X_4$, the shape is non-elliptical.

Table 3. Estimated parameters of Gaussian copula function

| Parameters | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| $rho_1$ | 0.42444 | 0.01695 | 25.042 | <2.00E-16 |
| $rho_2$ | 0.02134 | 0.02183 | 0.977 | 0.32836 |
| $rho_3$ | 0.06535 | 0.02213 | 2.953 | 0.00314 |
| $rho_4$ | -0.03431 | 0.02114 | -1.623 | 0.10466 |
| $rho_5$ | 0.0408 | 0.02130 | 1.915 | 0.05544 |
| $rho_6$ | 0.53365 | 0.01344 | 39.711 | <2.00E-16 |

Source: own calculations, based on FADN data

To allow for a different marginal distribution and non-elliptical shape of the 2-dimensional distribution, the Gaussian copula function was estimated with such parameter values as given in Table 3. The correlations from Table 3 show the fairly strong positive relation between yields of wheat and rape, and between prices of wheat and rape. All other correlations are very weak and not significant at a typical 5% significance level in most cases.

As mentioned in the introduction, the main aim of this paper was to investigate whether a copula function will outperform the multivariate normal distribution in modelling the joint distribution of crop plant yields and prices. For that purpose, the Voung test was used. Since this is a test relatively little known to the majority of agriculture economists, an example of a calculation is given below:

$$z_V = \frac{(-34702.76) - (-35179.8) - \frac{15-14}{2}\log(2268)}{\sqrt{2268 \cdot 0.5013}} = 14.03 \tag{6}$$

Comparing the $z_V$ statistic with quantiles of the standard normal distribution N(0, 1), we can see that the hypothesis of equidistance from the true model must be rejected on a arbitrarily low level of significance, i.e., p-value is below 2.00E-16. Therefore, it must be concluded that modelling joint distribution of crop plant yields and prices on the basis of a copula function is definitely a better choice than using the multivariate normal distribution.

Figures 4 and 5 show scatterplots for the samples generated from joint distribution of crop plant yields and prices based on a copula function and on the estimated multivariate normal distribution, respectively. It is clear that only the first one allows for the non-elliptical 2-dimensional distribution observed in the empirical data. It is a visual confirmation of the above tests, which show that the multivariate normal distribution is not suitable for modelling the joint distribution of crop plants yields and prices.

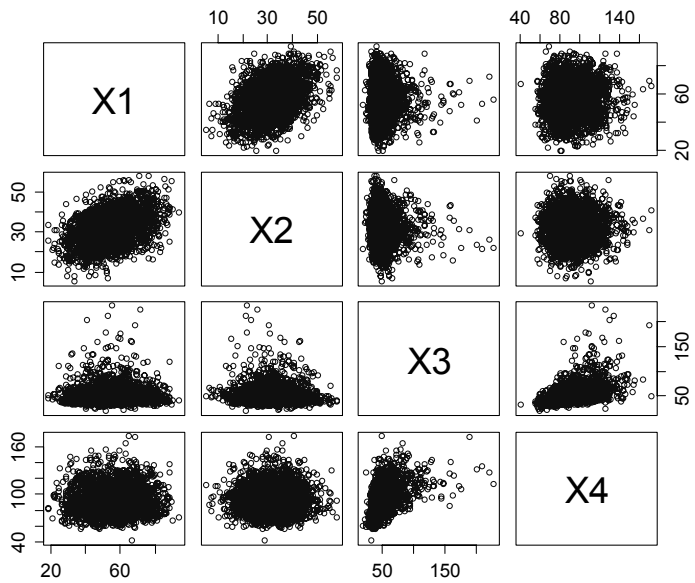


Fig. 4. Sample data generated with the model based on the estimated Gaussian copula function
Source: own calculations

Fig. 5. Sample data generated with the estimated multivariate normal distribution
Source: own calculations

The results so far indicate the clear advantage of using the copula based joint distribution, but to demonstrate how important it could be in practice to choose the right distribution, quantiles of income for a given crop structure were calculated.

Table 4. The relative discrepancies between empirical income quantiles for a given crop structure and the theoretical income quantiles (based on estimated join distributions)

| Probability | Empirical [PLN] | Copula f. | Normal distribution |
|---|---|---|---|
| | Crop structure - 10% winter wheat, 90% rape | | |
| 0.01 | 1290 | 4.4% | -4.3% |
| 0.02 | 1461 | 1.6% | -4.9% |
| 0.05 | 1727 | -1.4% | -4.2% |
| 0.10 | 1939 | -0.9% | -2.1% |
| 0.50 | 2820 | -0.5% | 1.4% |
| | Crop structure - 90% winter wheat, 10% rape | | |
| 0.01 | 1250 | -3.0% | -24.5% |
| 0.02 | 1357 | -1.0% | -17.7% |
| 0.05 | 1629 | -3.6% | -13.9% |
| 0.10 | 1808 | -1.4% | -7.0% |
| 0.50 | 2701 | -1.0% | 3.7% |

Source: own calculations, based on FADN data

It can be noted, on the basis of table 4, that for the {10% wheat, 90% rape} structure, both the joint distributions behave quite well, with the relative difference being less than 5%. But for the {90% wheat, 10% rape} structure, only the copula based distribution performs just as well as for the previous structure. The multivariate normal distribution gives differences of up to 25%. The reason for that could be the marginal distribution of wheat prices. The share of wheat in the first structure is too small for the wheat prices to be really of any importance when an inappropriate distribution is selected, but in the second case, when the share of wheat is so high, then choosing the inappropriate distribution clearly distorts the arguments which follow.

## Conclusions

The ability of incorporating different marginal distributions by a copula function is vital for joint modelling of crop plant yields and prices.

Joint distribution of crop plant yields and prices modelled with the use of a Gaussian copula function constitutes a significant improvement over the multivariate normal distribution, i.e., it has a significantly better fit to empirical data.

In the case of high-skew variables, such as the price of wheat, the Burr distribution has a significantly better fit than a log-normal distribution which is traditionally used to model the distribution of prices.

Using an inappropriate joint distribution of crop plants yields and prices results in the unreliable estimation of income distribution for the crop structures being analysed.

## References


Dutang C., Goulet V., Pigeon M. [2008]: actuar: An R Package for Actuarial Science. *Journal of Statistical Software*, vol. 25, no. 7, pp. 1-37. URL http://www.jstatsoft.org/v25/i07.

Hofert M., Kojadinovic I., Maechler M., Yan J., [2013]: copula: Multivariate Dependence with Copulas. R package version 0.999-7. URL http://CRAN.R-project.org/package=copula.

R Core Team [2013]: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schulte-Geers M., Berg E. [2011]: Modelling farm production risk with copulae instead of correlations. Paper prepared for presentation at the EAAE 2011 Congress Change and Uncertainty Challenges for Agriculture, Food and Natural Resources August 30 to September 2, Zurich.

Sklar A. [1959]: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, pp. 229-231.

Tadikamalla P. R. [1980]: A look at the Burr and related distributions. *International Statistical Review*, Vol. 48, Number 3, pp. 337–344.

Tejeda H.A., Goodwin B.K. [2008]: Modeling Crop prices through a Burr distribution and Analysis of Correlation between Crop Prices and Yields using a Copula method. Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Orlando, July 27-29.

Voung Q.H. [1989]: Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*,
Vol. 57, No. 2, pp. 307-333.

Zhu Y., Ghosh S.K., Goodwin B.K. [2008]: Modeling Dependence in the Design of Whole Farm Insurance Contract, |A Copula-Based Model Approach. Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Orlando, July 27-29.