BOGDAN WOLSKI
Kielce University of Technology
e-mail: bwolski@tu.kielce.pl

CEZARY TOŚ
Cracow University of Technology

Manuscript submitted 2017.09.19 - revised 2018.01.16,
initially accepted for publication 2018.02.16, published in March 2018

THE STATISTICAL MODELS FOR INTERPRETING THE RESULTS OF PARCELS AREA VERIFICATION IN INTEGRATED ADMINISTRATION AND CONTROL SYSTEM IACS

STATYSTYCZNE MODELE INTERPRETACJI WYNIKÓW WERYFIKACJI POWIERZCHNI DZIAŁEK W PROGRAMIE IACS

DOI: 10.30540/sae-2018-005

Abstract

A key element of the system IACS is the verification of the parcel area covered by direct subsidies. Control measurements are made by FOTO method, and in a small part by the direct inspection. Statistical methods are used in estimating the significance of differences. The results of such analysis are correct only when the empirical distributions are consistent with the theoretical ones. The problem of distribution adequacy is presented in the paper on the examples of three objects. The hypotheses about the possibility of using the commonly used distributions, and the appropriateness of the modification of the density curves were verified. By questioning the effectiveness of current methods of analysis, the authors point to the advantages of robust statistics. The cognitive effect of the analysis is to indicate the Laplace distribution as a statistical model of the analyzed differences. Research is concluded by proposal of post-control report that sums up relevant properties of the survey results.

Keywords: system IACS, verification of parcel area, robust statistics

Streszczenie

Kluczowym elementem programu IACS (Integrated Administration and Control System) jest weryfikacja powierzchni działek objętych dopłatami bezpośrednimi. Pomiary kontrolne wykonywane są metodą FOTO, a w niewielkiej części w ramach inspekcji terenowej. W ocenie istotności różnic zastosowanie znajdują metody statystyczne. Wyniki takich analiz są poprawne pod warunkiem zgodności rozkładów empirycznych z teoretycznymi. Problem adekwatności rozkładów zaprezentowano w artykule na przykładzie trzech obiektów. Zweryfikowano hipotezę o możliwości wykorzystania powszechnie stosowanych rozkładów oraz zbadano zasadność modyfikacji krzywych gęstości. Poddając w wątpliwość efektywność stosowanych obecnie metod analizy, Autorzy wskazują na zalety metod statystyki odpornościowej. Poznawczym efektem analizy jest wskazanie rozkładu Laplace'a jako statystycznego modelu analizowanych różnic. Konkluzję badań stanowi propozycja raportu pokontrolnego zawierającego istotne właściwości wyników pomiaru.

Słowa kluczowe: program IACS, weryfikacja powierzchni działek, statystyka odpornościowa

1. Introduction

ARiMR (Restructuring and Modernization of the Agriculture Agency) studies, reports and analysis of the control measurements of parcels covered by direct payments always contain several measures, namely: the number of reference parcels, the areas measured by farmers incorrectly, the percentage of

1. Wprowadzenie

W opracowaniach ARiMR raporty i analizy wyników pomiarów kontrolnych powierzchni działek objętych dopłatami bezpośrednimi standardowo zawierają kilka miar, a mianowicie: liczbę działek referencyjnych, których powierzchnie zostały przez rolników określone błędnie, procentowy udział tych

these parcels in the studied population, the mean or median observed of the surface differences [6-9]. In literature about the subject problem there are no advanced statistical studies. In statistical data analysis, such as testing hypotheses about the similarity of distributions, it is assumed, that the measurement results are random variables with a normal distribution. Such assumption without proper verification is not eligible. Investigation of numerical models quality worked out by photogrammetric methods shows, that errors of these models are better approximated by the Laplace's distribution. The histograms of empirical Digital Elevation Models (DEM) have a characteristic shape, they are centered near zero and they have so called "long tails" [1-3, 11]. For clarity of interpretation of the results, it is necessary to eliminate outliers data. Such effect can be achieved using the robust method [4].

A separate question to be clarified at the stage of data preparation for statistical analysis is the definition of a random variable. In practice, the elements of the set are calculated as absolute values in the differences between the reference parcel areas that were given by farmers and the results of the control survey. However, the subject of the analysis may also be relative values of these differences The problem is significant because, as experience shows, these variables are characterized by different distributions [3].

The problem of the properties of random variables tested and their corresponding methods of analysis have been investigated on the basis of the data obtained for three objects. Two specific problems of statistical interpretation were considered. One was to analyze the differences in the area of areas between farmers' declarations and FOTO control results, the second one was to compare the results of measurements made on the same objects at two year interval. An attempt was made to determine the theoretical distribution of the random variable best fitted to empirical data. It should be emphasized, that the theoretical distributions fitting to empirical data are of practical importance. On the one hand, they record the results of the measurements in a compact way, and on the other hand they enable the correct use of statistical tests. The result of the research is a proposition to take into account the statistical robust method in preparing IACS reports.

działek w badanej populacji, średnie lub mediany obserwowanych różnic powierzchni [6-9]. W literaturze przedmiotowego problemu brakuje zaawansowanych opracowań statystycznych. W statystycznych analizach danych, np. przy testowaniu hipotez o podobieństwie rozkładów, przyjmuje się, że wyniki pomiaru są zmiennymi losowymi o rozkładzie normalnym. Założenie takie bez odpowiedniej weryfikacji nie jest uprawnione. Badania jakości numerycznych modeli opracowanych metodą fotogrametryczną pokazują, że błędy tych modeli lepiej aproksymuje rozkład Laplace'a. Histogramy empirycznych modeli NMT mają charakterystyczny kształt, są skupione w pobliżu zera oraz posiadają tzw. „długie ogony" [1-3, 11]. Dla przejrzystości interpretacji wyników konieczne jest wyeliminowanie danych odstających. Taki efekt można uzyskać, stosując metodę statystyki odpornościowej (*robust method*) [4].

Oddzielną kwestią, którą należy uściślić na etapie przygotowania danych do analiz statystycznych, jest zdefiniowanie zmiennej losowej. W praktyce elementy zbioru obliczane są jako bezwzględne wartości różnic pomiędzy powierzchniami działek referencyjnych podanymi przez rolników oraz wynikami pomiaru kontrolnego. Jednak przedmiotem analizy mogą być również przedmiotowe różnice w postaci wartości względnych. Problem jest istotny, bowiem jak pokazują doświadczenia, wymienione zmienne charakteryzują się różnymi rozkładami [3].

Problem właściwości badanych zmiennych losowych i adekwatnych im metod analizy podjęto w prezentowanym tu artykule, na przykładzie danych pozyskanych dla trzech obiektów. Rozważono dwa szczegółowe zadania statystycznej interpretacji. Jedno dotyczyło analizy różnic powierzchni działek pomiędzy deklaracją rolników a wynikiem kontroli metodą FOTO, w drugim porównano rezultaty pomiarów wykonanych na tych samych obiektach w odstępie dwóch lat. Podjęto próbę określenia rozkładu teoretycznego badanej zmiennej losowej najlepiej dopasowanego do danych empirycznych. Warto podkreślić, że teoretyczne rozkłady dopasowane do danych empirycznych mają znaczenie praktyczne. Z jednej strony w zwarty sposób zapisują wyniki pomiarów, z drugiej umożliwiają poprawne stosowanie testów statystycznych. Efektem wykonanych badań jest propozycja konstrukcji raportu uwzględniająca metodę statystyki odpornościowej.

2. Survey of reference parcels areas by the FOTO method

The primary unit of the Land Parcel Identification System (LPIS) is the reference parcel. It is an agricultural land use, geographically defined and with an assigned identification number. Parcel areas are updated in the LPIS database based on orthophotomap on economic scope PEG and reference borders and GO. PEG is the area determined in the IACS system as the maximum land use area where crops may appear and are eligible for payment, while GO is the boundaries of land parcels updated according to actual use. Note that the definition of agricultural land eligible for direct payments differs from the definition of agricultural land in Land and Buildings Register.

Every year, 5% which is about 70,000 farms is controlled [14]. The measurement is done by the FOTO method because the field inspection of such a number of farms would be too expensive. The FOTO method utilises aerial photographs and high precision satellite images as well as field photos that complement and help identifying crops. Data from the LPIS database on the area are applied to orthophotomaps, then the vectorization of agricultural parcels is conducted. As a result, the boundaries of the parcels and the areas not eligible for payment are pre-identified. Photographs and other information from the field inspection are applied to the orthophotomap and are returned to the inspectors, who carry out the proper vectorization (Fig. 1). The areas of the reference parcels should be determined precisely, because it determines the amount of direct payments.

The orthophotomaps necessary for the implementation of the inspections are produced on the basis of aerial photographs or satellite imagery. The orthophotomap pixel dimensions are of 25 cm or 50 cm. Aerial images previously used by ARiMR were recorded in visible (VIS) and near infrared (NIR) range with a Root-Mean-Square-Error (RMSE) of ±0.5 m for resolution of 0.25 m and ±1.25 m for 0.5 m resolution. In the satellite imagery VIS and NIR spectral ranges are recorded. Resolutions of the imagery from the satellites used in the studies are: Quick Bird – 0.6 m, GeoEye – 0.5 m, IKONOS – 1.0 m, WorldView1 – 0.5 m. The final product is an orthophotomap in the scale of 1:5000 in Poland CS92 coordinate system.

2. Pomiar powierzchni działek referencyjnych metodą FOTO

Podstawową jednostką systemu LPIS (Land Parcel Identification System) jest działka referencyjna. Jest ona użytkiem rolnym, określonym pod względem geograficznym i z przyporządkowanym numerem identyfikacyjnym. Powierzchnie działek są aktualizowane w bazie danych LPIS w oparciu o ortofotomapę w zakresie ewidencyjno-gospodarczym (PEG) oraz granic odniesienia GO. PEG jest powierzchnią określaną w programie IACS jako maksymalny obszar, na którym może występować uprawa kwalifikowana do płatności, GO natomiast oznaczają granice działek ewidencyjnych uaktualnione według faktycznego stanu użytkowania. Zauważmy, że definicja gruntów rolnych kwalifikujących się do płatności bezpośrednich różni się od definicji użytków rolnych w EGiB.

Rocznie kontrolą objętych jest 5%, czyli ok. 70 000 gospodarstw [14]. Pomiar wykonuje się metodą FOTO, bowiem inspekcja terenowa takiej liczby gospodarstw byłaby zbyt kosztowna. Podstawę metody FOTO stanowią precyzyjne zobrazowania satelitarne bądź lotnicze oraz zdjęcia wykonane w terenie, które uzupełniają i pomagają w identyfikacji upraw. Dane z bazy LPIS dotyczące obszaru są nakładane na ortofotomapy, następnie prowadzona jest wektoryzacja działek rolnych. W jej wyniku wstępnie identyfikuje się granice działek rolnych oraz powierzchnie niekwalifikujące się do płatności. Zdjęcia oraz pozostałe informacje z inspekcji terenowej są nanoszone na ortofotomapę i wracają do inspektorów kameralnych, którzy przeprowadzają właściwą wektoryzację (rys. 1). Powierzchnia działki referencyjnej powinna być wyznaczona dokładnie, bowiem decyduje o wysokości dopłat bezpośrednich.

Ortofotomapy niezbędne do realizacji kontroli są wykonywane na podstawie zdjęć lotniczych bądź zobrazowań satelitarnych. Wielkość piksela ortofotomapy w skali terenowej wynosi 25 cm bądź 50 cm. Zdjęcia lotnicze dotychczas wykorzystywane przez ARiMR rejestrowane były w zakresach VIS (zakres widzialny) oraz NIR (bliska podczerwień) z błędem RMSE (Root-Mean-Square-Error) ±0,5 m dla rozdzielczości terenowej zobrazowania 0,25 m oraz ±1,25 m dla rozdzielczości 0,5 m. W zobrazowaniach satelitarnych rejestrowane są zakresy spektralne VIS+NIR. Rozdzielczości przykładowych zobrazowań z satelitów wykorzystywanych do opracowań to Quick Bird – 0,6 m, GeoEye – 0,5 m, IKONOS – 1,0 m, WorldView1 – 0,5 m. Produktem finalnym jest ortofotomapa w skali 1:5000 w układzie PL-1992.

Fig. 1. Comparison of vectors of reference and agricultural land use

Rys. 1. Porównanie wektorów działki referencyjnej oraz rolnej

According to the Specifications [13], the effect of ambiguity of border vectorization on parcel area is estimated by the tolerance T determined as the product of its circumference O and the width of the buffer zone B:

$$T = B \cdot O \quad (1)$$

The width of the buffer zone is determined arbitrarily by the ARiMR for the whole area, based on the results of the validation of the provided orthophotomap. The validation process is described in detail in the papers [2] and [5]. In case when ARMiR does not validate the orthophotomaps, an inspector calculates the width of the buffer zone as:

$$B = 1.5 \cdot R_t \quad (2)$$

where R_t is the resolution of the orthophotomap in m.

In the case where the difference between the declared area and the control's result exceeds the value of T, then the error of the farmer's declaration is declared. If the area is exceeded, the code DR13+ is given and, in case of underestimation, the code DR13-.

3. Research of the compatibility of declarations with the actual state

The presented research material is the result of the control of three objects in the West Pomerania Province. A total of 511 parcels in the year of 2014 were controlled and 365 parcels in 2016. The subject of the analysis was the number of plots of land for which T was exceeded and the percentage of incorrect declarations in two above mentioned periods. Tolerance T-values have been included in the analysis for the different widths of the buffer zones. In 2014, the buffer zone was 0.75 m and in 2016 it was 1.25 m. Consequently, the above data were analyzed in the second row based on the harmonized

Zgodnie ze Specyfikacjami [13] wpływ niejednoznaczności wektoryzacji granic na wielkość powierzchni działki szacuje się za pomocą tolerancji T wyznaczanej jako iloczyn jej obwodu O i szerokości strefy buforowej B:

$$T = B \cdot O \quad (1)$$

Szerokość strefy buforowej jest określana arbitralnie przez ARiMR dla całego obszaru na podstawie wyników walidacji dostarczonych ortofotomap. Proces walidacji opisano szczegółowo w [2] i [5]. W przypadku braku walidacji ortofotomap przez ARMiR wykonawca oblicza szerokość strefy buforowej jako

$$B = 1{,}5 \cdot R_t \quad (2)$$

gdzie R_t jest rozdzielczością ortofotomapy, m.

W przypadku gdy różnica pomiędzy powierzchnią deklarowaną a wynikiem kontroli przekracza wartość T, wówczas stwierdza się błąd deklaracji rolnika. W przypadku zawyżenia wielkości powierzchni działce nadawany jest kod DR13+, a w przypadku niedoszacowania powierzchni kod DR13-.

3. Badania poprawności zgodności deklaracji ze stanem faktycznym

Prezentowany materiał jest rezultatem kontroli trzech obiektów w województwie zachodniopomorskim. W roku 2014 skontrolowano łącznie 511 działek, a w 2016 roku 365 działek. Przedmiotem analizy były: liczba działek, dla których stwierdzono przekroczenie wartości T, oraz procentowy udział nieprawidłowych deklaracji w dwóch okresach. W analizach przy wyznaczaniu wartości tolerancji T uwzględniono różne szerokości stref buforowych. W 2014 roku strefa buforowa wynosiła 0,75 m, a w 2016 roku 1,25 m. Konsekwentnie w drugiej kolejności przeanalizowano wyżej wymienione wielkości przygotowane na podstawie

data. The standardization consisted of calculating new tolerances for a uniform buffer's width.

Compatibility of the farmer's declaration with the actual situation is best illustrated by the number of parcels for which codes DR13+ or DR13- have been assigned. The number of plots of land with significant discrepancies in declared area and actually used in 2014 and 2016 equals 126, corresponding to 14.4% of total data. Significant differences can be observed in individual years. In 2014, 18% of plots of land were declared wrongly (92 out of 511), and in 2016 only 9% (34 out of 365). The research question is to determine the significance of this difference in the aspect of statistical properties of data sets. The answer is ambiguous. Positive answer may indicate a better recognition of the system by farmers, but the decrease in the number of parcels with incorrectly declared area may also be the result of a different buffer zone acceptance. To verify this, a new width of the buffer zone same as in 2016 (1.25 m), was introduced for 2014 data, and new tolerances were calculated for each plot of land. With these assumptions, the number of parcels with incorrectly declared area has decreased from 92 to 80, from 18% to 16% of the total number of measurements. Let us note, that this result is practically identical to the value of 16.1% obtained for the West Pomerania Province in 2007 [9].

ujednoliconych danych, które polegało na obliczeniu nowych tolerancji dla jednolitej szerokości bufora.

Zgodność deklaracji rolników ze stanem faktycznym najlepiej obrazuje liczba działek, dla których przyporządkowano kody DR13+ lub DR13-. Liczba działek, w przypadku których stwierdzono istotne rozbieżności powierzchni deklarowanej i faktycznie użytkowanej w latach 2014 i 2016, wynosi łącznie 126, co odpowiada 14,4% ogólnej liczby danych. Znaczne różnice można zaobserwować w poszczególnych latach. W roku 2014 błędnie zadeklarowano 18% działek (92 działki spośród 511), a w 2016 jedynie 9% (34 spośród 365). Zagadnieniem badawczym jest określenie istotności tej różnicy w aspekcie statystycznych właściwości zbiorów danych. Odpowiedź nie jest jednoznaczna. Pozytywna może wskazywać na lepsze rozpoznanie problemu funkcjonowania systemu przez rolników, ale zmniejszenie się liczby działek, których powierzchnie zadeklarowano nieprawidłowo, może również wynikać z przyjęcia różnej strefy buforowej. Aby to sprawdzić, dla danych z 2014 roku wprowadzono nową szerokość strefy buforowej, taką jak w 2016 r., tj. 1,25 m, po czym dla każdej działki obliczono nowe tolerancje. Przy tych założeniach liczba działek o źle zadeklarowanej powierzchni zmniejszyła się z 92 do 80, czyli z 18% do 16% ogólnej liczby pomiarów. Odnotujmy, że wynik ten jest praktycznie identyczny z wartością 16,1%, jaki uzyskano dla województwa zachodniopomorskiego w 2007 r. [9].

Table 1. Number of parcels with DR13 code for different buffer zone width
Tabela 1. Liczba działek z kodami DR13 dla różnej szerokości strefy buforowej

	2016	2014	2014 (buffer width 1.25 m)	Difference between 2014 and 2016 (buffer width 1.25 m)
DR13+	18	51	44	7
DR13-	16	41	36	5
total	34	92	80	12

4. Properties of empirical data distribution

Figure 2 shows the histograms for three random variables, each slightly differentiating the declared and measured use area. These are:

1. Absolute differences in declared and measured area Δp.
2. Relative differences $\Delta p/p$.
3. Distances of boundaries (Author's designation), defined as mean distances between declared and measured boundaries d. It is a variable characterizing the relative differences:

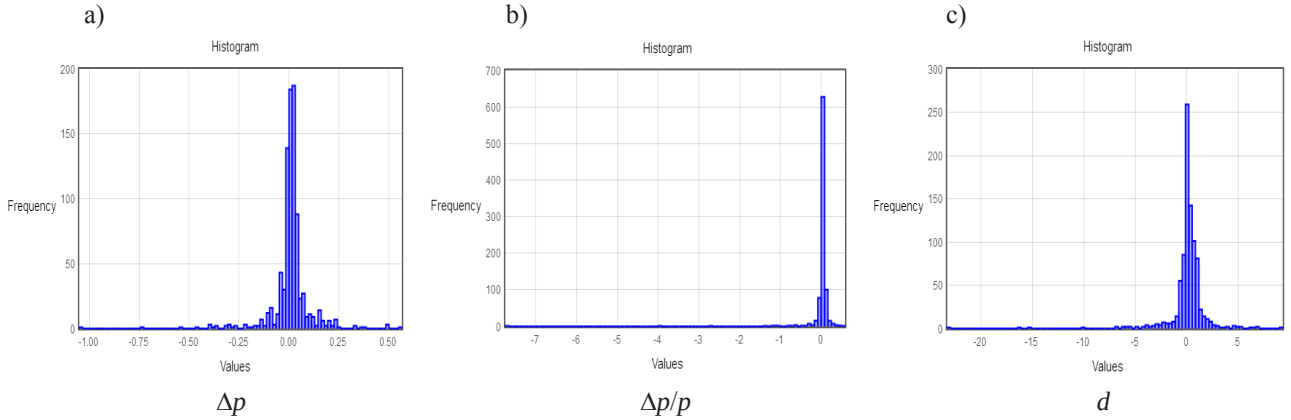
4. Właściwości rozkładu danych empirycznych

Na rysunku 2 przedstawiono histogramy wyznaczone dla trzech zmiennych losowych, z których każda w nieco inny sposób charakteryzuje różnice powierzchni deklarowanych i pomierzonych. Są to:

1. Bezwzględne różnice powierzchni deklarowanej i zmierzonej Δp.
2. Różnice względne $\Delta p/p$.
3. Odległość granic (określenie Autorów), wyznaczana jako średnia odległość pomiędzy granicami deklarowanymi i zmierzonymi d, jest zmienną charakteryzującą różnice względne:

$$d = \frac{\Delta p}{O} \qquad (3)$$

$$d = \frac{\Delta p}{O} \qquad (3)$$



Δp $\Delta p/p$ d

Fig. 2. Distribution histograms for the entire population of 876 parcels: a) absolute area differences, b) relative areas differences, c) average border distances

Rys. 2. Histogramy rozkładów dla całej populacji 876 działek: a) bezwzględne różnice powierzchni, b) względne różnice powierzchni, c) średnie odległości granic

The histograms for the analyzed variables Δp, $\Delta p/p$, d are characterized by "long tails" and a large focus near zero. On the basis of a visual assessment, the hypothesis, that the population model is a normal distribution, may be rejected. Objective evaluation was obtained by comparing the empirical distributions of the data for the years 2014 and 2016 for variables Δp, $\Delta p/p$, d with normal, normal-modified distributions [4] and Laplace distribution. The probability density functions in the above distributions are:

– normal

$$f_N(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \qquad (4)$$

where: x – random variable (Δp, $\Delta p/p$, or d), μ – mean value, σ – standard deviation;

– modified normal distribution

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\cdot NMAD}\exp\left(\frac{-(x-m)^2}{2\cdot NMAD^2}\right) \qquad (5)$$

where: m – median, $NMAD$ – normalized absolute deviation of median calculated as

$$NMAD = 1.4826\cdot Median\left(\left|x_i - m\right|\right) \qquad (6)$$

– Laplace' distribution

$$f_L(x) = \frac{1}{2b}\exp\left(\frac{-(x-m)^2}{b}\right) \qquad (7)$$

where b is a scale parameter calculated as

Kształty histogramów dla analizowanych zmiennych Δp, $\Delta p/p$, d charakteryzują się „długimi ogonami" i dużym skupieniem w pobliżu zera. Już na podstawie wizualnej oceny można odrzucić hipotezę, że modelem populacji jest rozkład normalny. Obiektywną ocenę uzyskano, porównując rozkłady empiryczne danych z lat 2014 i 2016 dla zmiennych Δp, $\Delta p/p$, d z rozkładami: normalnym, zmodyfikowanym normalnym [4] i rozkładem Laplace'a. Funkcje gęstości prawdopodobieństwa w tych rozkładach mają postać:

– normalny

$$f_N(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \qquad (4)$$

gdzie: x – zmienna losowa (Δp, $\Delta p/p$ lub d), μ – wartość średnia, σ – odchylenie standardowe;

– zmodyfikowany rozkład normalny

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\cdot NMAD}\exp\left(\frac{-(x-m)^2}{2\cdot NMAD^2}\right) \qquad (5)$$

gdzie: m – mediana, $NMAD$ – znormalizowane odchylenie bezwzględne mediany obliczane jako

$$NMAD = 1.4826\cdot Median\left(\left|x_i - m\right|\right) \qquad (6)$$

– rozkład Lapalace'a

$$f_L(x) = \frac{1}{2b}\exp\left(\frac{-(x-m)^2}{b}\right) \qquad (7)$$

gdzie b to parametr skali obliczany jako

$$b = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - m|$$

the Laplace distribution variance equals to:

$$Var = 2 \cdot b^2 \quad (8)$$

Matching empirical data to theoretical distribution is presented in the form of quantile plots Q-Q, depicting the relationship between empirical and theoretical quantiles. As a measure of the accuracy of fitting empirical data into the theoretical distribution, Root-Mean-Square-Deviation (*RMSD*) was calculated using the formula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} y_i}{n}} \quad (9)$$

where: y_i – deviations between the empirical histogram and assumed probability density function, n – population size.

The thesis on the similarity between empirical and theoretical distribution was verified using the Shapiro-Wilk test and the λ-Kolmogorov-Smirnov test. After identification of outlying observations by the Grubbs test, elements of the robust statistics were introduced.

$$b = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - m|$$

wariancja w tym rozkładzie wynosi:

$$Var = 2 \cdot b^2 \quad (8)$$

Dopasowanie danych empirycznych w rozkład teoretyczny przedstawiono w postaci wykresów kwantylowych Q-Q, obrazujących zależność kwantyli teoretycznych od empirycznych. Jako miarę dokładności wpasowania danych empirycznych w rozkład teoretyczny przyjęto *RMSD* (Root-Mean-Square-Deviation) obliczaną według wzoru:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} y_i}{n}} \quad (9)$$

gdzie: y_i – odchyłki pomiędzy histogramem empirycznym i przyjętą funkcją gęstości prawdopodobieństwa, n – liczebność populacji.

Tezę o podobieństwie rozkładu empirycznego do teoretycznego zweryfikowano za pomocą testu Shapiro-Wilka oraz testu λ-Kołmogorowa-Smirnowa. Elementy statystyki odpornościowej wprowadzono po identyfikacji obserwacji odstających za pomocą testu Grubbsa.

Table 2. Parameters of normal distribution for empirical data

Tabela 2. Parametry rozkładu normalnego dla danych empirycznych

	Random variable		
	Δp	$\Delta p/p$	d [m]
2014	sample size 508		
Mean	-4.3 m^2	-0.038	-0.09 m
Standard deviation	992.20 m^2	0.412	1.91 m
Kurtosis	14.38	250.84	52.02
RMSD	0.531	0.841	0.567
Shapiro-Wilk statistics W	0.713	0.174	0.683
W – critical 5% significance level	0.994		
λ-Kołmogorowa	4.41	6.91	4.64
λ max – 95% significance level	1.36		
2016	sample size 363		
Mean	66.0 m^2	-0.006	0.070 m
Standard deviation	937.05 m^2	0.175	1.454 m
Kurtosis	52.78	168.71	48.78
RMSD	0.482	0.713	0.497
Shapiro-Wilk statistics W	0.627	0.282	0.625
W – critical 5% significance level	0.992		
λ-Kołmogorowa	3.02	4.59	2.93
λ max – 95% significance level	1.36		

The results of the measurement were first analyzed for outlier observations. Using the Grubbs test, three observations were rejected for the Δp, $\Delta p/p$, in the 2014 data set and two observations in the 2016 data set. The results of fitting the remaining empirical data into the normal distribution are summarized in Table 2 and Figure 3.

Wyniki pomiaru w pierwszej kolejności poddano analizie pod kątem występowania obserwacji odstających. Stosując test Grubbsa, odrzucono trzy obserwacje dla zmiennych Δp, $\Delta p/p$, d w zbiorze danych z 2014 roku i dwie obserwacje w zbiorze danych z roku 2016. Wyniki wpasowania pozostałych danych empirycznych w rozkład normalny zestawiono w tabeli 2 i na rysunku 3.

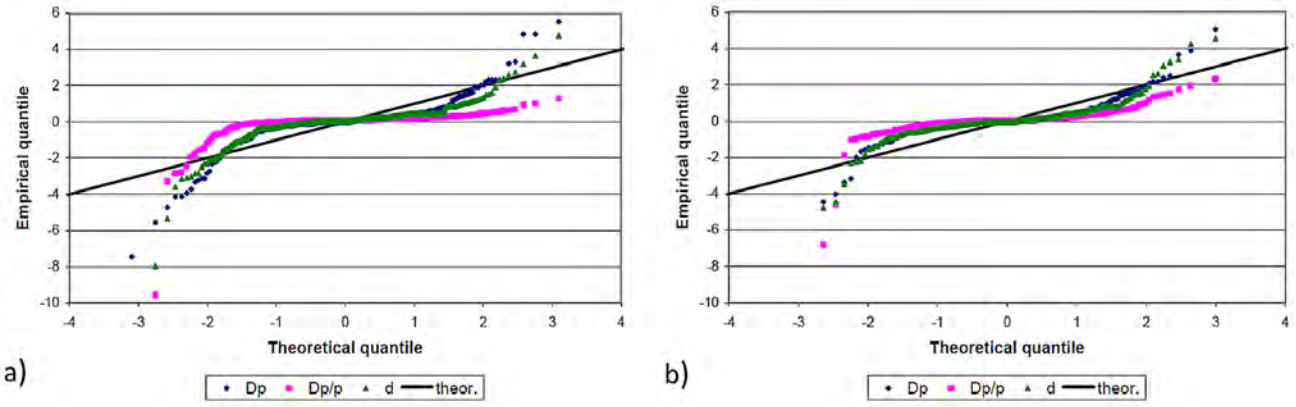


Fig. 3. Q-Q plots fitting empirical data into normal distribution: a) data from 2014, b) data from 2016
Rys. 3. Wykresy Q-Q wpasowania danych empirycznych w rozkład normalny: a) dane z 2014 roku, b) dane z 2016 roku

The results of the analysis shown in Figure 3a, Figure 3b and in Table 2 indicate that Δp, $\Delta p/p$, d are not random variables with normal distribution. Statistical values of the Shapiro-Wilk test in each case are well below the critical level in both populations. The normal distribution hypothesis should also be rejected in case of the λ-Kolmogorov-Smirnov test comparing the cumulative curve with the distribution function. The distribution of variables Δp, $\Delta p/p$, d differs significantly, although they are all leptokurtic and therefore very slim. The largest concentration around the mean is the distribution of the variable $\Delta p/p$ (high kurtosis). This distribution is also the most common deviation from the normal distribution (large RMSD values, λ and very low W values). The other two distributions have similar parameters, but the distribution of the variable d differs from the normal distribution by more than the distribution of the variable Δp. The results of fitting empirical data into the modified normal distribution and Laplace distribution are shown in Figure 4 and 5 and in Tables 3 and 4.

Wyniki analizy pokazane na rysunkach 3a i 3b oraz w tabeli 2 wskazują, że Δp, $\Delta p/p$, d nie są zmiennymi losowymi o rozkładzie normalnym. W obu badanych populacjach wartości statystyki W testu Shapiro-Wilka są w każdym przypadku znacznie mniejsze niż poziom krytyczny. Hipotezę o rozkładzie normalnym odrzucić także należy w przypadku testu λ-Kołmogorowa-Smirnowa porównującego krzywą kumulacyjną z dystrybuantą. Rozkłady zmiennych Δp, $\Delta p/p$, d znacznie się różnią, jakkolwiek wszystkie mają charakter leptokurtyczny, a więc bardzo wysmukły. Największe skupienie wokół wartości średniej ma rozkład zmiennej $\Delta p/p$ (duże wartości kurtozy). Rozkład ten najbardziej też odstaje od rozkładu normalnego (duże wartości RMSD, λ i bardzo małe wartości statystyki W). Pozostałe dwa rozkłady mają zbliżone do siebie parametry, z tym że rozkład zmiennej d różni się od rozkładu normalnego minimalnie bardziej niż rozkład zmiennej Δp. Wyniki wpasowania danych empirycznych w zmodyfikowany rozkład normalny i rozkład Laplace'a przedstawiono na rysunkach 4 i 5 oraz w tabelach 3 i 4.

Table 3. Parameters of modified normal distribution for empirical data

Tabela 3. Parametry zmodyfikowanego rozkładu normalnego dla danych empirycznych

	Random variable		
	Δp	$\Delta p/p$	d [m]
2014	sample size 508		
Median	0.0 m^2	0.0	0.0 m
NMAD	317.62 m^2	0.030	0.56 m
RMSD	2.372	7.292	2.170
λ-Kolmogorov	1.84	1.65	1.67
λ max – 95% significance level 1.36			
2016	sample size 363		
Median	0.0 m^2	0.0	0.0 m
NMAD	296.52 m^2	0.022	0.46 m
RMSD	1.777	3.819	1.754
λ-Kolmogorov	1.67	1.38	1.57
λ max – 95% significance level 1.36			

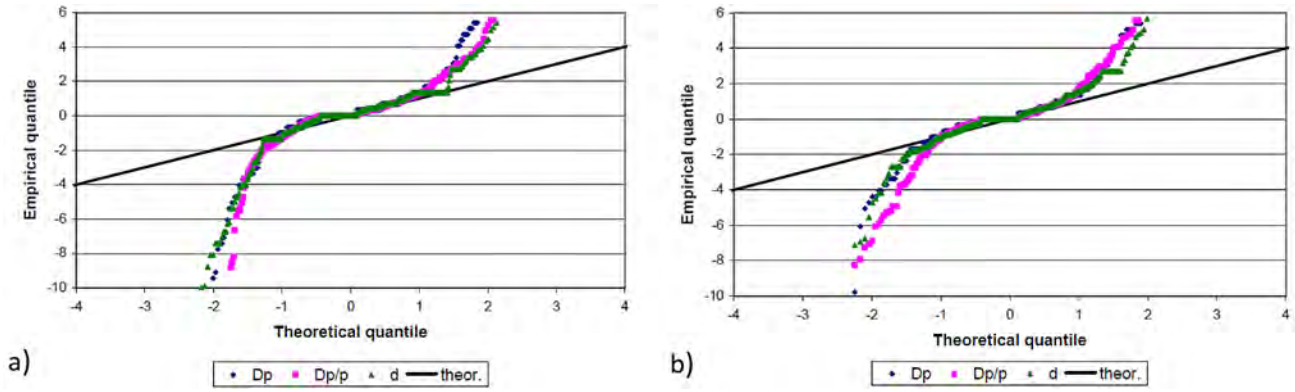


Fig. 4. Q-Q plots fitting empirical data into normal distribution modified: data from 2014, b) data from 2016

Rys. 4. Wykresy Q-Q wpasowania danych empirycznych w rozkład normalny zmodyfikowany: a) dane z 2014 roku, b) dane z 2016 roku

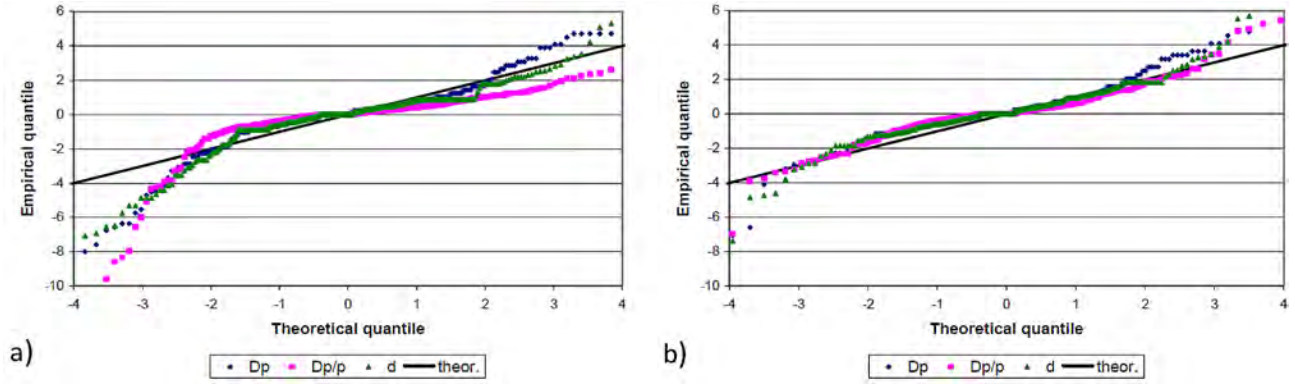


Fig. 5. Q-Q plots fitting empirical data intoLaplace's distribution: a) data from 2014, b) data from 2016

Rys. 5. Wykresy Q-Q wpasowania danych empirycznych w rozkład Laplace'a: a) dane z 2014 roku, b) dane z 2016 roku

Table 4. Parameters of Laplace distribution for empirical data
Tabela 4. Parametry rozkładu Laplace'a dla danych empirycznych

	Random variable		
	Δp	$\Delta p/p$	d [m]
2014	sample size 508		
Median	0.0 m^2	0.0	0.0 m
B	0.049 m^2	0.080	0.847 m
$\sqrt{\text{Variance}}$	0.0690 m^2	0.1132	1.1984 m
RMSD	0.795	2.362	0.897
λ-Kolmogorov	1.93	2.48	2.02
λ max – 95% significance level 1.36			
2016	sample size 363		
Median	0.0 m^2	0.0	0.0 m
B	0.044 m^2	0.048	0.678 m
$\sqrt{\text{Variance}}$	0.0621 m^2	0.0676	0.9584 m
RMSD	0.508	1.130	0.551
λ-Kolmogorov	1.80	1.87	1.56
λ max – 95% significance level 1.36			

From the λ-Kolmogorov-Smirnov test, it is clear that with a probability greater than 95%, the hypothesis that the empirical data have a normal and Laplace's distribution is to be rejected. At the same time, it is concluded from those two distributions, that empirical data better approximates Laplace distribution. However, the RMSD values of all variables for Laplace distribution are higher than for normal distribution. The modified normal distribution differs significantly from the empirical data. Research confirms the significant influence of both the random factor and the off-balance factors. Data sets can not therefore be qualified as coming from the same population. Data histograms have features that predispose them to the Laplace distribution or modified normal distribution (large focus around 0 and "long tails"). However, even for them the size of outliers in measurements are too large, as can be seen in Q-Q plots. In this situation, it seems appropriate to use a robust approach by filtering diverging data, which is supposed to belong to another population. It was found that the identification of diverging observations by the Grubbs test is not correct because the populations do not have a normal distribution and show large variances. For this reason, the alternative often used in measuring problems is to reject observations that do not fall within the range of $[u_{\text{mean}} - 2\sigma_u,\ u_{\text{mean}} + 2\sigma_u]$. For such reduced sets match tests to normal, modified normal and Laplace distributions were performed (Table 5).

The *RMSD* values and the λ-Kolmogorov test result indicate that there are no reasons for rejecting the hypothesis that the model of the corrected set

Z testu λ-Kołmogorowa-Smirnowa wynika, że z prawdopodobieństwem powyżej 95% należy odrzucić hipotezę, że dane empiryczne mają rozkład zmodyfikowany normalny i Laplace'a. Jednocześnie na podstawie dwóch rozpatrywanych powyżej rozkładów stwierdza się, że dane empiryczne lepiej przybliżają rozkład Laplace'a. Wartości *RMSD* wszystkich zmiennych dla rozkładu Laplace'a są jednak większe niż w przypadku rozkładu normalnego. Zmodyfikowany rozkład normalny odbiega zaś znacznie od danych empirycznych. Badania potwierdzają istotny wpływ czynnika losowego, jak i czynników generujących wartości odstające. Zbiorów danych nie można kwalifikować jako pochodzących z tej samej populacji. Histogramy danych posiadają cechy predysponujące do rozkładu Laplace'a lub zmodyfikowanego rozkładu normalnego (duże skupienie wokół 0 i długie „ogony"), nawet dla nich wielkości obserwacji odstających są zbyt duże, co można zaobserwować na wykresach Q-Q. W tej sytuacji właściwe wydaje się zastosowanie podejścia odpornościowego polegającego na odfiltrowaniu danych odstających, które z założenia należą do innej populacji. Stwierdzono, że identyfikacja obserwacji odstających testem Grubbsa nie jest poprawna, gdyż populacje nie mają rozkładu normalnego i wykazują duże wariancje. Alternatywą często stosowaną w zagadnieniach pomiarowych jest odrzucenie obserwacji niemieszczących się w przedziale $[u_{śr} - 2\sigma_u, u_{śr} + 2\sigma_u]$. Dla tak wydzielonych zbiorów przeprowadzono testy zgodności z rozkładami: normalnym, zmodyfikowanym normalnym i Laplace'a (tabela 5).

Wartości *RMSD* oraz wynik testu λ-Kołmogorowa wskazują, że nie ma podstaw dla odrzucenia hipotezy,

Table 5. Results of fitting distributions into empirical data
Tabela 5. Wyniki wpasowania rozkładów w dane empiryczne

	Normal distribution		Modified normal distribution		Laplace distribution	
	2014	2016	2014	2016	2014	2016
Δp variable	32 observations discarded in 2014, and 17 in 2016					
RMSD	0.326	0.295	0.728	0.722	0.336	0.315
λ-Kolmogorov	2.82	2.76	2.16	1.14	1.24	1.04
λ max – 95% significance level 1.36						
$\Delta p/p$ variable	13 observations discarded in 2014, and 11 in 2016					
RMSD	0.515	0.544	1.814	1.566	0.687	0.429
λ-Kolmogorov	4.05	2.86	1.78	1.62	1.84	1.39
λ max – 95% significance level 1.36						
d variable	28 observations discarded in 2014, and 17 in 2016					
RMSD	0.305	0.201	0.610	0.633	0.354	0.273
λ-Kolmogorov	2.80	1.82	1.96	1.73	1.50	1.40
λ max – 95% significance level 1.36						

of variables Δp is a Laplace distribution. It should be highlighted that the relatively positive results of λ-Kolmogorov test follow from the analysis of cumulative value. Accumulation of frequencies smoothes the irregularities characteristic for the empirical distribution plots.

5. Proposal for a post-control report

Identification of the appropriate data distribution allows for the extension of the post-control report. In addition to the percentages given in 4 we can add the following data (Table 6, Figure 6).

- Parameters of the theoretical distribution for the corrected dataset $[u_{mean} - 2\sigma_u, u_{mean} + 2\sigma_u]$.
- Characteristics of the outliers by mean y_{mean} and variance interval $[y_{max}, y_{min}]$, separately for left- and right-hand values.
- The result of comparison of both populations by Kolomogorov-Smirnov test without elimination of outliers (in practice this point would not be obligatory).

że modelem skorygowanego zbioru zmiennej Δp jest rozkład Laplace'a. Relatywnie pozytywne oceny, jakie uzyskuje się przy teście λ-Kołmogorowa, wynikają stąd, że przedmiotem analizy są wartości skumulowane. Kumulowanie częstości wygładza nieregularności charakterystyczne dla wykresu rozkładów empirycznych.

5. Propozycja raportu pokontrolnego

Identyfikacja odpowiedniego rozkładu danych umożliwia rozszerzenie raportu pokontrolnego. Autorzy sugerują, by oprócz wartości procentowych (jak w punkcie 4) podawać także:

- parametry rozkładu teoretycznego dla skorygowanego zbioru danych $[u_{śr} - 2\sigma_u, u_{śr} + 2\sigma_u]$,
- charakterystyki wartości odskakujących za pomocą wartości przeciętnej $y_{śr}$ oraz przedziału – zmienności $[y_{max}, y_{min}]$ oddzielnie dla wartości lewo- i prawostronnych,
- wynik porównania obu populacji testem Kołmogorowa-Smirnowa bez eliminacji obserwacji odstających (w praktyce ten punkt nie byłby obligatoryjny).

Table 6. Sample of post-control report
Tabela 6. Przykład raportu

Statistics	2014	2016
Distribution of variable Dp	Laplace	Laplace
Median Dp	0 m^2	0 m^2
Mean Dp	41 m^2	72 m^2
$\sqrt{\text{Variance}}$	393 m^2	411 m^2
Probability of difference greater than 1a and 10a	0.697 0.028	0.705 0.031
Part of outliers observations „+" (mean area) in % and m^2	3.1% 2812 m^2	2.7% 2655 m^2
Part of outliers observations „–" (mean area) in % and m^2	3.5% 3466 m^2	1.9% 3943 m^2
λ-Kolmogorov	0.54, which means that the probability of rejecting the hypothesis of distribution similarity is 0.06	

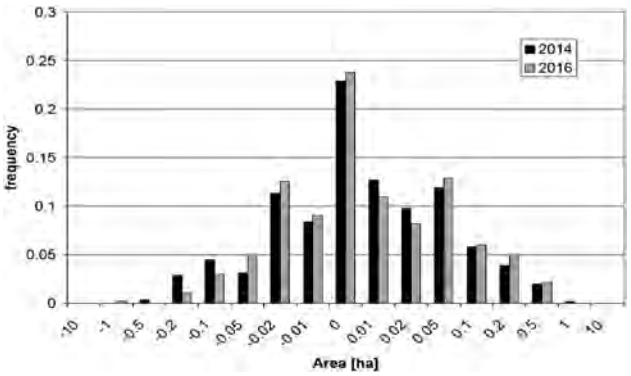


Fig. 6. Histograms of Δp variable for data from 2014 and 2016 years
Rys. 6. Histogramy zmiennej Δp dla danych z lat 2014 i 2016

6. Conclusions

- Statistical methods may be used to interpret the results of control measurements of parcel areas covered by direct payments. The differences between declared and measured areas cannot be modeled by using the normal distribution.
- In the interpretation of the results of control measurements, it is appropriate to use robust statistics methods. After removing outliers observations, the empirical population is best modeled by Laplace distribution.
- In the analyzed example, the best compatibility with the theoretical distribution was obtained for the variable Δp (difference of the declared and measured area) and the variable d determined from the quotient of the parcel area error and its circumference. An adverse effect was obtained with a variable defined as the relative difference of the area $\Delta p/p$.
- The effect of changing the width of the buffer and hence the tolerance is negligible. This is due to the number of parcels that have been incorrectly declared.
- Carrying out the control procedure produce a "didactic" effect. In the following years areas of references parcel areas are determined with better accuracy.
- Identification of data distribution allows for the extension of the scope of post-control report. Further standardization of data processing and preparation of post-control report requires more extensive research.

6. Wnioski

- Przy interpretacji wyników pomiarów kontrolnych powierzchni działek objętych dopłatami bezpośrednimi można stosować metody statystyczne. Różnic pomiędzy powierzchniami deklarowanymi i pomierzonymi nie można modelować za pomocą rozkładu normalnego.
- W procedurze interpretacji wyników pomiarów kontrolnych zasadne jest zastosowanie metod statystyki odpornościowej. Po usunięciu obserwacji odstających populację empiryczną najlepiej modeluje rozkład Laplace'a.
- W analizowanym przykładzie najlepszą zgodność z rozkładem teoretycznym uzyskano dla zmiennej Δp (różnic powierzchni deklarowanej i zmierzonej) oraz zmiennej d wyznaczonej z ilorazu błędu powierzchni działki i jej obwodu. Niekorzystny rezultat uzyskano w przypadku zmiennej zdefiniowanej jako względna różnica powierzchni $\Delta p/p$.
- Wpływ zmiany szerokości bufora, a tym samym tolerancji jest nieznaczny, wynika to z liczby działek, które zostały zidentyfikowane jako błędnie zadeklarowane.
- Przeprowadzenie kontroli ma efekt „uczący". W następnych latach wielkości powierzchni działek referencyjnych są określane dokładniej.
- Identyfikacja rozkładu danych umożliwia rozszerzenie obecnego zakresu raportu pokontrolnego. Dalsza standaryzacja procedury opracowania danych i redakcji końcowego raportu wymaga obszerniejszych badań.

References

[1] Darnel A.R., Nicholas J., Tate N.J., Chris Brunsdon C., 2008, *Improving user assessment of error implications in digital elevation models*, Computers, Environment and Urban Systems 32.
[2] Hejmanowska B., Drzewiecki W., Kulesza Ł., 2008, *Zagadnienie jakości numerycznych modeli terenu*, Archiwum Fotogrametrii, Kartografii i Teledetekcji, ISSN 2083-2214, vol. 18a, s. 163-175.

[3] Hejmanowska B., 2013, *Zastosowanie rozkładu Laplace'a do określania niepewności danych przestrzennych na przykładzie NMT i systemu IACS*, s. 150, Wydawnictwa AGH, Kraków 2013, ISBN 978-83-7464-649-9.
[4] Höhle J., Höhle M., 2009, *Accuracy assessment of digital elevation models by means of robust statistical methods*, ISPRS Journal of Photogrammetry and Remote Sensing 64, s. 398-406.
[5] Kramarczyk P., Hejmanowska B., Dąbrowski J. 2012, *Walidacja odbiorników GNSS dla potrzeb kontroli wielkości powierzchni działek rolnych w systemie dopłat bezpośrednich dla rolnictwa IACS: wytyczne, aplikacja, problemy*, Wydawnictwo Państwowej Wyższej Szkoły Techniczno-Ekonomicznej, Jarosław, ISBN 978-83-88139-52-9.
[6] Orlińska J., Wasilewska Z., *System odniesień przestrzennych LPIS komponent infrastruktury danych przestrzennych, Infrastruktura Danych Przestrzennych w Polsce i Europie* – Seminarium AR, Wrocław, 1-3XII 2004 http://www.gislab.up.wroc.pl/download/ Wasilewska_www_gislab_ar_wroc_pl.pdf?GISLabSessionID=shqv1polkduq6mqi3u66dmhg26.
[7] Wężyk P., Szostak M., Tompalski P., *Comparison of the accuracy of the "PHOTO" check method with automatic analysis based on ALS data for direct control of subsidy payment*, Archiwum Fotogrametrii, Kartografii i Teledetekcji, Vol. 20, 2009, s. 445-456, ISBN 978-83-61-576-10-5.
[8] Pośnik R., *Wykorzystanie różnych źródeł informacji na potrzeby zarządzania kryzysowego – System Identyfikacji Działek Rolnych (LPIS)*, Agencja Restrukturyzacji i Modernizacji Rolnictwa, Warszawaa 2014, www.2014.5zywiolow.pl/wp-content/uploads/2012/04/3-3-robert-posnik.pdf.
[9] Pradziadowicz M., *Kontrola gospodarstw rolnych w ramach programu rozwoju obszarów wiejskich 2007-2013*, Folia Pomer. Univ. Technol. Stetin., Oeconomica 2014, 313(76)3, s. 105-114.
[10] Lipiec A., *System Identyfikacji Działek Rolnych (LPIS) i jego powiązanie z EGIB* [w]: *Acta Scientifica Academiae Ostroviensis*, Wyższa Szkoła Biznesu i Przedsiębiorczości w Ostrowcu Świętokrzyskim 2011, nr 35-36, s. 167-172.
[11] Zandbergen P.A., 2008, *Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy*, Transactions in GIS,12(1), s. 103-130.
[12] Europejski Trybunał Obrachunkowy, *Sprawozdanie specjalne: system identyfikacji działek rolnych – użyteczne narzędzie do określania kwalifikowalności gruntów rolnych wymagające udoskonaleń w zakresie zarządzania*, Luksemburg: Urząd Publikacji Unii Europejskiej 2016.
[13] Specyfikacja Istotnych Warunków Zamówienia na prowadzenie i optymalizację LPIS w zakresie opracowania ortofotomapy na podstawie scen satelitarnych [ORTO_SAT_2016-2018]). http://www.arimr.gov.pl/aktualnosci/artykuly/dzp-2610-22016.html.
[14] Informacja o wyborze najkorzystniejszej oferty, numer postępowania DZP-2610-2/2016. http://www.arimr.gov.pl/uploads/media/Informacja_o_wyborze_najkorzystniejszej_oferty_ba34fe.pdf
[15] Załącznik nr 2 do umowy: *Instrukcja realizacji kontroli w zakresie kwalifikowalności powierzchni wersja 2.0,* ARiMR, Warszawa 2016.

Acknowledgments:
This work was supported by Kielce University of Technology, Grant No. 05.0.09.00/2.01.01.01.0014 MNSP.IKGG.14.001

Podziękowania:
Praca była finansowana przez Politechnikę Świętokrzyską, grant nr 05.0.09.00/2.01.01.01.0014 MNSP.IKGG.14.001