Barometr Regionalny
Tom 14 nr 4

# An Attempt of Delimitation of Polish Metropolitan Areas by Discriminant Analysis

**Jadwiga Zaród**
West Pomeranian University of Technology Szczecin, Poland

**Abstract**

*The purpose of this article is a division of Poland into areas with different economic and social conditions and the identification of the areas which may aspire to the name of metropolis. Initial division of Polish territorial units was made on the basis of statistical data on the labor market, wages, public utilities, education, health care, environment, culture, industry and construction. These data were subjected to standardization and were verified as to whether represent a normal distribution. Then the discriminatory power of the variables were examined and parameters of the linear discriminant function were estimated. The highest average value of the discriminant function indicates the area most developed in terms of the examined features. Territorial units that belong to this complex are metropolises or aspire to being them. For each area, the classification functions were estimated and on their basis the final division was made. The allocation to the sub-region was mainly determined by variables such as density of population, unemployment rate, average flat surface and the number of entrepreneurs. In the first group are territorial units that belong already to the metropolitan areas. Subsequent numbers indicate areas with increasingly less economic and social development.*

**Keywords:** data standardization, discriminant analysis, classification function, metropolitan areas

## Introduction

The act on spatial planning and development of 25 March 2003 includes a definition of a metropolitan area.[1] According to the act, a metropolitan area is a metropolitan settlement system with the surrounding and functionally-related areas. The Polish Metropolitan Union Council expanded the definition with "a settlement complex inhabited by over 500 thousand individuals that includes international cooperation institutions."[2] Each metropolitan area is characterized by:

- high quality of services, institutions and material equipment,
- high potential for technological, economic, social, political and cultural innovation,
- high competitiveness in production and specialized services (including scientific research and cultural services) on a national and international basis,
- strong internal bonds of economic, social and institutional cooperation,
- intensive connection with other national and foreign metropolises, made possible via good communication with them, and
- uniqueness and specificity of the site, as well as its attractiveness, not only on national but also international scale (Świetlik and Lubiatowski 2004).

With the exception of a few large metropolitan centers, it is difficult in Poland to find examples of fully developed metropolises with decision-making functions on a global scale.

The aim of this paper is to determine, with discriminant analysis, areas covering sub-regions with similar conditions of economic and social progress. The determined discriminant function

1. See: Ustawa z dnia 27 marca 2003 r. o planowaniu i zagospodarowaniu przestrzennym. DzU z 2003 r. nr 80 poz. 717.

2. See: Zaktualizowana Koncepcja Przestrzennego Zagospodarowania Kraju. Rządowe Centrum Studiów Strategicznych, Warszawa, październik 2005, page 93.

© 2016 by Wyższa Szkoła Zarządzania i Administracji w Zamościu
All Rights Reserved

value will indicate sub-regions aspiring to the title of metropolitan areas, and order the determined centers from least to most developed. The numerous conducted studies — among others by Jałowiecki (2000), Gierańczyk (2009), Smętkowski, Jałowiecki and Gorzelak (2009), Danielewicz and Turała (2011) or Młodak (2012) — on delimitation of Polish metropolitan areas analyzed mainly large cities and their surrounding areas. The scope of this research envelops all the sub-regions of the country, described with statistical data regarding the job market, salaries, municipal management, education, health and environmental care, culture, industry and construction in 2014.

Discriminant analysis, the main research method, will allow for a division of Poland into areas differing from each other in terms of selected characteristics. Next, classification methods will determine which of the created areas a given region should be assigned to, using the variables with the highest discriminatory power.

## 1 Division of Poland into areas

Each sub-region of Poland has been described with selected statistical data (*Rocznik Statystyczny…* 2015) called diagnostic variables. These variables are presented in table 1. Based on the diagnostic data from 2014, the sub-regions were initially divided into 7 areas (groups). The number of groups $k$ was determined based on the formula:

$$k = 1 + 3{,}322 \log n, \tag{1}$$

where $n$ (i.e., number of sub-regions) = 72.

The input data was standardized to make the analyses independent from individual variables' measurement units with the formula

$$z_{ij} = \frac{x_{ij} - \bar{x}}{s}, \tag{2}$$

where:
$x_{ij}$ — the value of $j$-th variable for $i$-th sub-region,
$\bar{x}$ — the mean value of a given variable,
$s$ — standard deviation (Zeliaś 2000).

**Tab. 1.** Diagnostic variables

| Variable | Variable description | Unit |
|---|---|---|
| $x_1$ | number of people | individuals |
| $x_2$ | population density | individuals per km$^2$ |
| $x_3$ | working-age population | % |
| $x_4$ | unemployment rate | % |
| $x_5$ | average monthly gross salary | PLN |
| $x_6$ | average usable area of 1 apartment | m$^2$ |
| $x_7$ | average usable area of an apartment per person | m$^2$ |
| $x_8$ | population benefiting from the water supply system | % |
| $x_9$ | population benefiting from the sewage system | % |
| $x_{10}$ | population benefiting from the gas supply system | % |
| $x_{11}$ | gross enrollment ratio for general high schools | % |
| $x_{12}$ | gross enrollment ratio for postsecondary schools | % |
| $x_{13}$ | number of students per 1 computer with Internet access in secondary schools | individuals |
| $x_{14}$ | number of students per 1 computer with Internet access in high schools | individuals |
| $x_{15}$ | number of people per 1 health care facility | individuals |
| $x_{16}$ | number of library subscribers per 1000 of population | individuals |
| $x_{17}$ | number of people per 1 place in theaters | individuals |
| $x_{18}$ | population with wastewater treatment plants' support | % |
| $x_{19}$ | number of national economy entities entered into the National Business Registry | – |

The $x_4$ variable is a destimulant, it was turned into a stimulant by multiplying its value by minus one. Discriminant analysis assumes that data (presented as variables) represents a sample from a multivariate normal distribution. After formation of abundance distributions, as well as performance of chi-square and Kolmogorov-Smirnov tests in order to assess normality of distribution, the number of variables was reduced to: $x_2$, $x_3$, $x_4$, $x_6$, $x_7$, $x_{13}$, $x_{14}$, $x_{15}$, $x_{16}$, $x_{18}$, $x_{19}$. Test results are included in table 2. The probabilities (except for the $x_2$ and $x_{19}$ variables) are higher than the $\alpha = 0{,}05$ level of significance, which gives no grounds for discarding the distribution's compatibility hypothesis. According to the chi-square test, the $x_2$ (population density) and $x_{19}$ (number of national economy entities) variable does not have a normal distribution, which is not directly proven by the Kolmogorov-Smirnov test. Because of these variables' significant importance in delimitation of metropolitan areas, they were retained for further analyses.

Further limitation of the variables' list resulted in study of discrimination power via the Wilks' lambda and $F$ tests. The analyses' outcomes are presented in table 3. The Wilks' Lambda statistic values are an assessment of a model's discrimination power after introduction of a given variable. The lower the value of that statistic, the higher the model's discrimination power after introduction of the variable. The value of the $F$ statistic is related to the individual contribution of each input variable and indicated the order of its introduction into the model. The highest contribution to discrimination, that the highest values of the $F$ statistic indicate, of various areas is attributed to the variables $x_2$, $x_4$, $x_6$, and $x_{19}$. The critical $p$-level close to 0 verifies the hypothesis that all the variables are relevant in a model explaining the diversity of sub-regions' clusters. The highest tolerance value equal to 0,8888 and $R$-squared equal to 0,1112 for the $x_{14}$ variable means that 88,88% of information introduced by that variable is not duplicated by other variables already present in the model.

**Tab. 2.** Normal distribution tests

| | Chi-square test | | Kołmogorov-Smirnov test | |
|---|---|---|---|---|
| **Variable** | ***H*** | ***p*** | ***D*** | ***p*** |
| $x_2$ | 12,910 | 0,005 | 0,137 | $< 0{,}1$ |
| $x_3$ | 1,678 | 0,642 | 0,044 | $> 0{,}2$ |
| $x_4$ | 2,073 | 0,557 | 0,034 | $> 0{,}2$ |
| $x_6$ | 2,587 | 0,463 | 0,034 | $> 0{,}2$ |
| $x_7$ | 2,410 | 0,492 | 0,045 | $> 0{,}2$ |
| $x_{13}$ | 3,677 | 0,159 | 0,043 | $> 0{,}2$ |
| $x_{14}$ | 1,830 | 0,660 | 0,055 | $> 0{,}2$ |
| $x_{15}$ | 1,162 | 0,559 | 0,042 | $> 0{,}2$ |
| $x_{16}$ | 5,218 | 0,156 | 0,097 | $> 0{,}2$ |
| $x_{18}$ | 0,216 | 0,898 | 0,018 | $> 0{,}2$ |
| $x_{19}$ | 18,638 | 0,008 | 0,147 | $< 0{,}1$ |

[In the journal European practice of number notation is followed—for example, 36 333,33 (European style) = 36 333.33 (Canadian style) = 36,333.33 (US and British style).—Ed.]

**Tab. 3.** Characteristics of variables in the model

| Variable | Wilks' Lambda | ***F*** | ***p*** | Tolerance | 1-Tolerance (***R*-squared**) |
|---|---|---|---|---|---|
| $x_2$ | 0,0114 | 46,4705 | $< 0{,}0001$ | 0,3922 | 0,6078 |
| $x_4$ | 0,0043 | 11,5374 | $< 0{,}0001$ | 0,4085 | 0,5915 |
| $x_6$ | 0,0037 | 8,4587 | $< 0{,}0001$ | 0,5165 | 0,4835 |
| $x_7$ | 0,0034 | 6,8874 | $< 0{,}0001$ | 0,3409 | 0,6591 |
| $x_{13}$ | 0,0033 | 6,4572 | $< 0{,}0001$ | 0,6956 | 0,3044 |
| $x_{14}$ | 0,0026 | 2,9805 | 0,0132 | 0,8888 | 0,1112 |
| $x_{18}$ | 0,0033 | 6,4804 | $< 0{,}0001$ | 0,7615 | 0,2385 |
| $x_{19}$ | 0,0034 | 7,0433 | $< 0{,}0001$ | 0,3981 | 0,6019 |

After assessing which variables differentiate (discriminate) areas most prominently, the parameters of linear discriminant functions were estimated:

$$F = A^T Z, \tag{3}$$

where:
$F = [f_{li}]$ — matrix of discriminant functions, where $f_{li}$ is the value of the $l$-th discriminant variable in the $i$-th sub-region,
$A = [a_{lj}]$ — matrix of discriminant variables' coefficients, where $a_{lj}$ is a coefficient located by the $l$-th discriminant variable and the $j$-th input (diagnostic) variable,
$Z = [z_{ji}]$ — standardized observation matrix, where $z_{ji}$ is the value of the standardized $j$-th variable in the $i$-th sub-region (Krzyśko 1990).

The maximum number of calculated functions is equal to the number of groups minus one (which is 6). Discriminant function in the form of

$$F = 1{,}4227x_2 + 0{,}2121x_4 + 0{,}1473x_6 + 0{,}0004x_7 + 0{,}1247x_{13} + 0{,}0580x_{14} + 0{,}0927x_{18} - 0{,}9586x_{19}$$

explains 79,74% of intergroup variance, has the lowest value of the Wilks' Lambda test (0,0019), and that is why it will be the basis for further analyses. The $x_2$, $x_{19}$ and $x_4$ variables have the greatest influence on the formation of this discriminant function. Standardized coefficients were used for assessment of individual variables' influence on discriminant functions' formation. These coefficients can be used for calculation of canonical values (discriminant functions' values) for each case, and of average values for each area, as well as for ordering the determined centers (Zawadzki and Babis 1996). The highest average value of discriminant function indicated the area best developed in terms of the studied characteristics. The territorial units that belong to that center are metropolises or aspire to be one. With the increase of an areas' number, their socio-economic development level decreases. Table 4 presents the calculated mean values for individual areas. In III, IV, V and VI areas, the mean canonical values are closer together, which proves a considerable similarity of the sub-regions they include. Area I significantly separates itself from the other centers.

**Tab. 4.** The average values of the discriminant function

| Area | Mean values canonical |
|---|---|
| I | 12,0408 |
| II | 0,2562 |
| III | −1,6605 |
| IV | −1,6877 |
| V | −1,6999 |
| VI | −1,7477 |
| VII | −2,4843 |

## 2 Classification of sub-regions

Classification methods assign a given object to one of pre-made groups, based on the variables with the highest discriminatory power. When initiating the classification of sub-regions, probability a priori proportional to the groups' magnitude was selected, for the large variance of economic and social conditions in Poland indicates that the formed areas will not be equally numerous. Next, each area was assigned a classification function in the form of

$$K_r = c_{r0} + c_{r1}x_1 + c_{r2}x_2 + \cdots + c_{rj}x_j \qquad r = 1, 2, \ldots, k \quad j = 1, 2, \ldots, m, \tag{4}$$

where:
$K_r$ — $r$-th classification variable (for the $r$-th group of sub-regions),
$c_{r0}$ — the constant for the $r$-th group,
$c_{rj}$ — coefficients of variables with significant discriminatory power,
$x_j$ — observed (standardized) values for the $j$-th variable (Krzyśko 1990).

The values of all classification functions were calculated for each sub-region. A given sub-region was assigned to an area for which it has the highest classification value. The accuracy value of initial classification and suggested changes based on classification functions are presented in table 5. Over 65% of regions were properly qualified for different areas. The result for region I was flawless (100%). The lowest percentage of correct classifications was found in group V (40%). After introduction of changes according to the specified classification, the classification functions were reassessed and their values were calculated for each sub-region. The research was redone until 100% correctness of classification was obtained.

**Tab. 5.** Initial classification

| Area | Correctness of classification (%) | Number of sub-regions in particular areas | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | I | II | III | IV | V | VI | VII |
| I | 100,00 | 8 | | | | | | |
| II | 87,50 | | 6 | 1 | | | | 1 |
| III | 60,00 | | 1 | 3 | 1 | 2 | 3 | |
| IV | 45,45 | | 1 | | 8 | 1 | | 2 |
| V | 40,00 | | | 1 | 2 | 4 | 1 | 2 |
| VI | 70,00 | | | 1 | 2 | | 10 | 1 |
| VII | 66,67 | | | | | 1 | 2 | 7 |
| *Average* | *65,15* | *8* | *8* | *6* | *13* | *8* | *16* | *13* |

*Note:* rows — initial classification, columns — division based on classification functions

Table 6 presents classification functions' variable coefficients based on which sub-regions were ultimately divided into areas. The higher the absolute value of the coefficients present by the variables, the higher the influence of those variables on the creation of classification functions and the classification of a given sub-region to a correct area. In areas I, III, IV, V, VI and VII, population density had a significant influence on the ultimate division of sub-regions. Classification depended on unemployment rate in four centers (I, II, VI and VII). The average area of an apartment largely determined the division in areas III and IV. The number of economic entities also influences the classification in centers I, III and IV.

**Tab. 6.** Coefficients of variables classification functions

| Variable | Area | | | | | | |
|---|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI | VII |
| $x_2$ | 57,836 | −2,9141 | −10,3330 | −6,9757 | −8,6957 | −6,3404 | −9,5772 |
| $x_4$ | 3,275 | 4,2490 | −3,4709 | 2,5135 | 1,8585 | −1,8623 | −4,2234 |
| $x_6$ | −2,8507 | −1,4364 | 6,3792 | 4,5852 | −1,4408 | −1,0474 | −1,9940 |
| $x_7$ | 0,3366 | −2,2141 | 1,5212 | −2,5805 | −1,9995 | 1,1969 | 2,5368 |
| $x_{13}$ | 1,8511 | 0,0409 | 2,6575 | 0,3173 | −3,3761 | −0,1686 | −0,4941 |
| $x_{14}$ | −0,7473 | 0,2419 | 2,5890 | −0,2361 | −0,0825 | −0,7550 | 0,7913 |
| $x_{18}$ | 1,8646 | 1,3424 | 2,7341 | −0,5683 | −1,1151 | −2,7639 | 1,9860 |
| $x_{19}$ | −14,3934 | 0,5225 | 4,5877 | 2,7564 | 1,4795 | 0,9370 | 2,0204 |
| Constant | −74,7535 | −4,9245 | −12,8893 | −5,8879 | −6,8654 | −4,1739 | −5,8661 |

Table 7 presents the ultimate classification of sub-regions to individual areas. Whereas table 8 contains the results of each sub-region's classification to the correct area. In the first area are territorial units that belong already to the metropolitan areas. While the sub-regions in the second area aspire to become a metropolis. Sub-regions in area I and II were marked on the figure 1.

**Tab. 7.** Ultimate classification

| Area | Correctness of classification (%) | Number of sub-regions in particular areas | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | I | II | III | IV | V | VI | VII |
| I | 100,00 | 8 | | | | | | |
| II | 100,00 | | 8 | | | | | |
| III | 100,00 | | | 4 | | | | |
| IV | 100,00 | | | | 14 | | | |
| V | 100,00 | | | | | 6 | | |
| VI | 100,00 | | | | | | 18 | |
| VII | 100,00 | | | | | | | 14 |
| *Average* | *100,00* | *8* | *8* | *4* | *14* | *6* | *18* | *14* |

*Note:* rows — initial classification, columns — division based on classification functions

**Tab. 8.** Classification of sub-regions to areas

| Area | Sub-regions |
|---|---|
| I | Warszawa (city), Łódź (city), Kraków (city), katowicki, Poznań (city), Wrocław (city), trójmiejski, Szczecin (city) |
| II | tyski, gliwicki, rybnicki, bydgosko-toruński, sosnowiecki, legnicko-głogowski, warszawski zachodni, poznański |
| III | warszawski wschodni, gdański, krakowski, wrocławski |
| IV | rzeszowski, oświęcimski, nowosądecki, nowotarski, leszczyński, kaliski, tarnowski, opolski, gorzowski, pilski, bielski, koniński, częstochowski, bytomski |
| V | lubelski, tarnobrzeski, krośnieński, przemyski, skierniewicki, puławski |
| VI | grudziądzki, chojnicki, starogardzki, świecki, łódzki, zielonogórski, piotrkowski, łomżyński, ciechanowski, ostrołęcki, siedlecki, płocki, suwalski, kielecki, chełmsko-zamojski, sandomiersko-jędrzejowski, bialski, sieradzki |
| VII | chojnicki, inowrocławski, białostocki, włocławski, wałbrzyski, olsztyński, radomski, nyski, ełcki, słupski, szczeciński, koszaliński, szczecinecko-pyrzycki, elbląski |

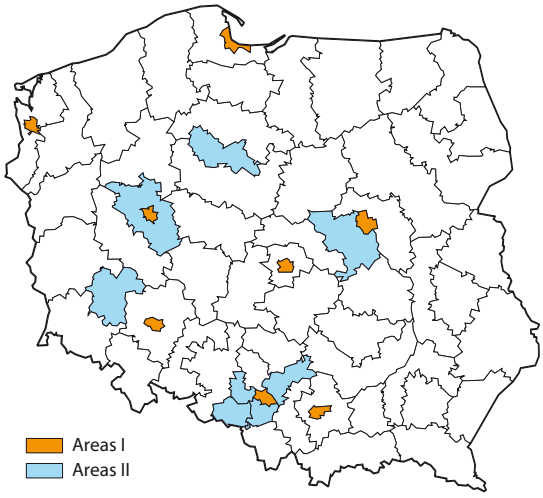


**Fig. 1.** Sub-regions I and II areas

## 3 Characteristics of areas

The most important characteristics of the singled out areas were presented on figures 2–5. The analysis concerns variables (more specifically, their values in 2014) that had a significant influence on the classification in several areas.

Area I included territorial units with very high population densities (from 1 355 individuals/km$^2$ in Szczecin to 3 355 individuals/km$^2$ in Warszawa). High population density could also be found in the gliwicki sub-region (543 individuals/km$^2$) included in area II. Population density in areas III, IV and V was comparable. The lowest population density was noted in area VII, more specifically in the szczecinecko-pyrzycki sub-region.

Unemployment rate rose with the region's number. In area I, the unemployment rate did not exceed 10% (the highest in Szczecin—9,3%), and only amounted to 3,1% in Poznań. The highest unemployment was noted in area VII, and the unemployment rate there varied from 13,3% in the białostocki sub-region to 22,7% in the ełcki sub-region.

The average apartment area rose from area I to area III, at which point it started decreasing. The lowest apartment area was at the disposal of the inhabitants of Łódz (average apartment area—53,8 m$^2$). On the other hand, the highest average apartment area belonged to the people in area III (from 83 m$^2$ in the warszawski wschodni sub-region to 94,4 m$^2$ in the krakowski sub-region).

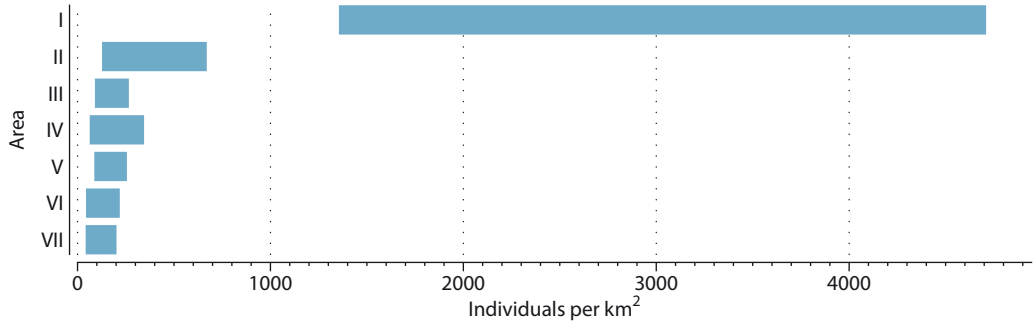


**Fig. 2.** Population density

*Note:* Placement and width of the bar show minimum (left edge) and maximum (right edge) values of sub-regions in particular area)

*Source:* Own work based on (*Rocznik Statystyczny…* 2015)

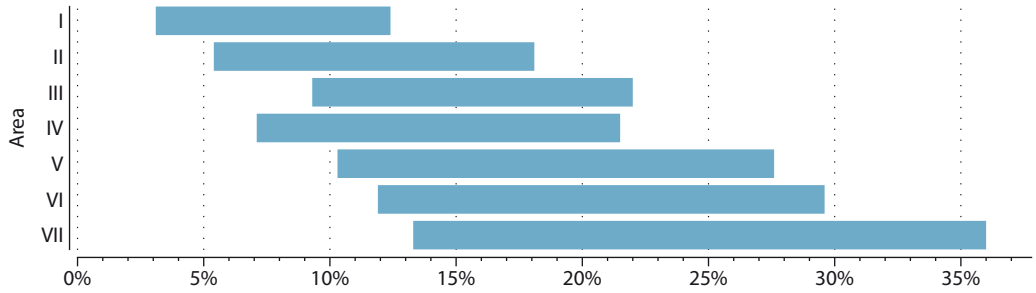


**Fig. 3.** Unemployment rate

*Source:* Own work based on (*Rocznik Statystyczny…* 2015)

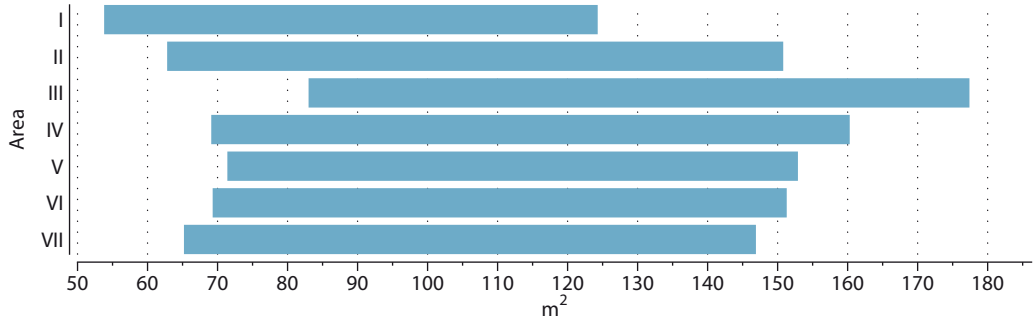


**Fig. 4.** Average apartment area

*Source:* Own work based on (*Rocznik Statystyczny…* 2015)

In Warsaw, the number of national economic entities entered into the National Business Registry in 2014 amounted to 371 476. A high number of economic entities was noted in the warszawski zachodni sub-region (109 326) and in the bydgosko-toruński sub-region (89 574). The lowest number of economic entities could be found in the suwalski sub-region (19 612), the ełcki sub-region (23 182) and the przemyski sub-region (25 967).

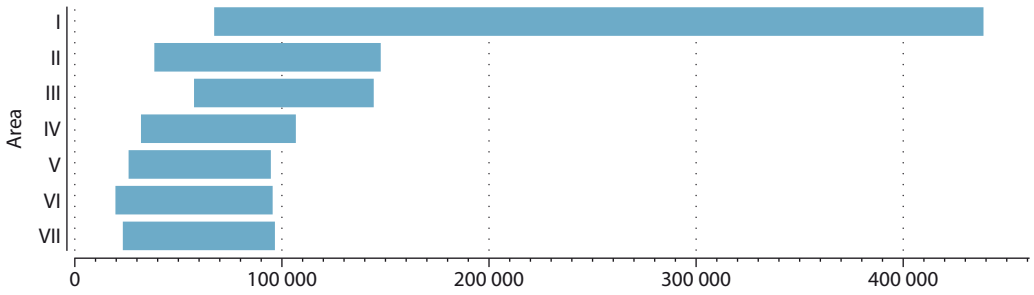


**Fig. 5.** National economic entities entered into the National Business Registry

*Source:* Own work based on (*Rocznik Statystyczny…* 2015)

## Conclusions

Assessment of discriminant functions was accomplished with the help of diagnostic variables with high discriminatory power. The variables with high influence on the formation of a given discriminant function's value included: population density, the number of national economic entities entered into the National Business Registry and unemployment rate. The highest average value of discriminant function indicated the area best developed in terms of the studied characteristics — a metropolitan area. Subsequent areas mean centers with increasingly lower economic and social development. The discriminatory function therefore allowed for a division of Poland into areas differing from each other in terms of selected variables, and ordering the selected centers from best to least developed. The division's accuracy was confirmed by Wilks' lambda test's low coefficient (0,0019). On the other hand, classification functions allowed for assigning individual sub-regions to determined areas. The variables that mainly determined the classification included: population density, unemployment rate, apartment area and the number of national economic entities entered into the National Business Registry. Population density in Poland is very diversified and varies in sub-regions (not cities) from 42 to 543 individuals per km$^2$. The lowest unemployment rate (3,1%) occurs in the area I and the highest in the ełcki sub-region (22,7%). Average apartment area falls within the limits of 53,8–94,4 m$^2$. The highest amount of registered economic entities can be found in Warszawa (371 476) and the lowest in the suwalski sub-region (19 612). Area I (Metropolitan) has the highest density of population, the lowest unemployment rate and the highest number of entities of the national economy.

It can be therefore concluded that a discriminant analysis can be used as a supporting tool for division of Poland into areas with different developmental capabilities and for separation of metropolitan areas.

## References

Danielewicz, J., and M. Turała. 2011. "Delimitacja obszarów metropolitalnych jako podstawa wdrażania metropolitan governance." *Acta Universitatis Lodziensis Folia Oeconomica* (258):109–121.

Gierańczyk, W. 2009. "Warunki życia ludności bydgosko-toruńskiego obszaru metropolitalnego." In *Aglomeracje miejskie w Polsce na przełomie XX i XXI wieku. Problemy rozwoju, przekształceń strukturalnych i funkcjonowania. Zbiór rozpraw*, edited by W. Maik, 231–246. Bydgoszcz: Wydawnictwo Uczelniane Wyższej Szkoły Gospodarki.

Jałowiecki, B. 2000. *Społeczna przestrzeń metropolii*. Warszawa: "Scholar."

Krzyśko, M. 1990. *Analiza dyskryminacyjna*. 2nd ed., *Statystyka Matematyczna*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Młodak, A. 2012. "Statystyka metropolii polskich. Problemy i perspektywy." *Studia Regionalne i Lokalne* (2):20–38.

*Rocznik Statystyczny Województw 2015*. 2015. Warszawa: Główny Urząd Statystyczny.

Smętkowski, M., B. Jałowiecki, and G. Gorzelak. 2009. *Obszary metropolitalne w polsce: problemy rozwojowe i delimitacja* (*Diagnoza problemów rozwoju obszarów metropolitalnych i rekomendacja delimitacji obszarów metropolitalnych w Polsce*). Edited by G. Gorzelak, *Raporty i analizy EUROREG*. Warszawa: Centrum Europejskich Studiów Regionalnych i Lokalnych EUROREG.

Stawasz, D. ed. 2004. *Ekonomiczno-organizacyjne uwarunkowania rozwoju regionu — teoria i praktyka*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Świetlik, M., and A. Lubiatowski. 2004. "Plany metropolii — jak je opracować? Cele, problematyka, tryb." *Wspólnota* (13):34–37.

Zawadzki, J., and H. Babis. 1996. *Zastosowanie analizy dyskryminacyjnej do oceny kondycji finansowej przedsiębiorstw, Zeszyty Naukowe Uniwersytetu Szczecińskiego*. Szczecin: Uniwersytet Szczeciński.

Zeliaś, A. 2000. *Metody statystyczne*. Warszawa: Polskie Wydawnictwo Ekonomiczne.