New paradigm in statistics and population census quality

Elżbieta Gołata, elzbieta.golata@ue.poznan.pl
Poznań University of Economics
Poland

Abstract
The paper refers to shift in methods applied in statistics: from traditional censuses through sample survey to administrative data, as a new paradigm in statistical surveys. The research is conducted with respect to population census. Special attention is put to assessing quality of 'new generation' population census which use multiple data sources. Register-based censuses in conjunction with sample survey are considered as well as opportunities and statistical challenges of the future.
Despite long tradition and well-developed research methodology, censuses do not provide 'perfect' results. Census data quality is discussed with respect to criteria resulting from its aim, applied census method, number of data sources used for evaluation and types of errors. First of all, census may be burdened with non-random: coverage and content errors. Due to different methods for conducting censuses, also random errors and nonresponse are analyzed. The paper aims to propose an integrated approach to assess quality of population census. In particular, the quality assessment of census data refers to Polish experiences, both for traditional (2002 census) and combined method (2011 census).
Discussion includes also harmonization problems, danger of divergent results and estimates, in terms of several data sources.
Keywords:
quality of population census; register-based census; random and non-random errors; coverage and content errors; census coverage survey

Introduction

Population census is not only the oldest research, best-known, well-formed in terms of methodology, but also a research, which is widely regarded as the most reliable data source. As methods of conducting census and especially of data collecting have changed incredibly over last decades, it seems appropriately important to address the issue of quality assessment of population census. The purpose of this paper is to discuss the quality of information derived from the 2011 population census in Poland. Special attention is given to comparison of census data accuracy in view of traditional versus register-based approach.

Register-based censuses have already been conducted from 70. Invaluable in this respect is the experience of the Nordic countries [11]. But the 2010 round brought a methodological shift in the way of doing censuses in many countries. The register-based approach and mixed method greatly expanded. These methods were applied also by countries whose population is much larger than in Scandinavian countries and with little experience in the use of administrative data in official statistics.

Poland is a country experiencing transformation in economy and social life. Nowadays the state is celebrating 10 years of membership in the European Union, with all the challenges and benefits, among them consequences resulting from adjustment of regulations in official statistics. Social assessment of the census and attention of the scientific community are diverse in nature. They are both full recognition and criticism. However, change in the approach to obtain information by public statistics involves at least a few methodological issues. Reference to them, in some double sense, is shown in the title of this paper.

Shift in paradigm

Paradigm [gr. parádeigma 'pattern'], as introduced by Thomas Kuhn is a set of concepts and theories forming foundations of the natural science: "Universally recognized scientific achievements that, for a time, provide model problems and solutions for a community of practitioners"[23]. The change applies to the nature of statistics, understood as a process of obtaining information which is the basis of analysis and inference. And it is not just the data collecting, but the whole process of statistical survey [18].

We are definitely in the process of paradigm shift in official statistics. Over a hundred years ago, if the information about the population were needed, it meant the census needed to be done. There was no other way of getting these data; census was the only way to get them[1]. Then, at the beginning of XX century, people thought of studying some representatives of the entire population. The representative method has been developed by the works of Jerzy Neyman, Karl Pearson and Sir Ronald Fisher. So a dozen years ago, when the information about the population was needed, it meant a sample survey had to be done. Nobody considered doing a census more often than once every 10 years. It was too expensive. But

[1] Other data sources, like books of the parish, or other administrative records were also used. J. Graunt in *Natural and Political Observations Made upon the Bills of Mortality* (1662) used the mortality rolls in London to construct first life tables.

nowadays a survey would not be carried out immediately, first one would rather look on registers, administrative records and other information that are available from different sources. This means a paradigm in statistics, new approach to obtain information [22].

It is beyond the scope of this paper to discuss the four concepts for statistics known as classical or error statistics, Bayesian statistics, Likelihood-based and Akaikean-Information Criterion-based statistics. Controversies in the foundation of statistics concern issues that have been debated for years without resolution [2].

Modern Information and Communication Technology provided an incredible amount of information from sources inaccessible so far, but of different quality and different properties. This is actually something not accidental at all. Big data are available in reality, but need evaluation, and integration. Therefore data editing, evaluation and integration is coming more and more relevant, also in the context of official statistics. This involves reorganization in statistics [20][21]. A number of problems could be discussed here. The most significant challenges include Small Area Estimation, Data Integration, GIS, different sources for statistical frames, e-surveys, e-Questionnaires, variety of methods for data dissemination.

Shift in population censuses methodology

The 2010 round of population censuses showed that many countries experienced transition in census methodology. Before all, change in data sources need to be mentioned. Instead of classical enumeration, data are extracted from administrative records. Instead of traditional field operations, Internet transfer is applied. UNSD carried out two surveys on how countries were implementing their national censuses. First survey was conducted between May 2009 and January 2010 and aimed at collecting information on methodologies used and implementation of modern technology during different phases of the census operation [15]. The second survey was undertaken in July 2011 with a follow-up with non-responding countries in mid-2012. The second survey had the objective of collecting information on the lessons learned from the 2010 round of population and housing censuses [16].

In the field of census methodology, according to data presented, majority (56%) of countries applied the traditional method, but the percentage of countries using register-based approach increased by 100%, accounting to 14.5% of the surveyed countries (tab. 1). However, if complex method would be also taken into account, this proportion rises from 18% in censuses

round 2000 to 40% in 2010 round [15]. In the scope of modern technology implemented in censuses, Internet transfer of information should be underlined. Internet was used primarily to increase the number of responses, minimize refusals to participate in the study, or to obtain information directly from respondents.

Table 1. UNECE countries by Census method for 2000 and 2010 round

Method	Census round	
	2000-2002	2010-2014
Traditional	40 (80%)	31 (56%)
Register-based	4 (8%)	8 (15%)
Mixed method	6 (12%)	14 (25%)

Source: UN 2012

2011 population census in Poland was conducted by applying mixed method with the use of administrative records (full survey - short form), supplemented by information from Internet self-enumeration. Additionally a sample survey (long form) was performed on approximately 20% of randomly selected dwellings. In preparation for the 2011 census metadata about 300 administrative registers were collected and analysed. All variables in those systems were rated with regards to the possibility of obtaining information on population, housing and buildings, in line with the recommendations and classifications of the UN Statistical Office (UND) and Eurostat [14]. As a result of detailed analysis 28 registers were selected. Among them as a priority the following systems should be mentioned: Universal Electronic System for Registration of Population, Social Security System; the Health Insurance System, Land and buildings, Register of Territorial Division of the Country, data from the State Fund for Rehabilitation of Persons with Disabilities. Properly structured and divided into strata information collected from administration sources were also used in creating frame for the census sample survey.

To assess quality of the census different methods may be used. The classification of sources of errors indicates coverage and content errors [1]. Among evaluation methods census coverage survey and different methods of demographic analysis are most popular [6]. As concerns coverage assessment, it is common to have a net census undercount as the number of omissions usually exceeds the number of duplications. But considering the change in census methodology, besides non-random, also random errors should be discussed. As the census used data from different sources, including, registers, self-enumeration, and sample results,

which were integrated in one data base, evaluation should be comprehensive and clear description of estimation technique provided.

Because there are different aspects of census evaluation, there is no one universal criterion of its quality. According to UNECE 2013 survey on national practices in the 2010 round of population and housing censuses, there was no full agreement to criteria for success [16]. Majority of the countries pointed „Overall user and stakeholder support", as the main criterion for evaluation. Among other criteria, one could find: public support, improved outputs, cost saving, improved response/participation rates, improved coverage rates, staff expertise, software etc. There was also no consensus, in view of the methodology used. For countries, that used traditional census: user and stakeholder support, improved coverage rates, government support, were the most important criteria. Countries that followed the combined approach, pointed out: improved response/participation rates, user and stakeholder support, cost savings. Cost savings, government support, improved outputs, were the most important to countries that conducted register-based censuses.

Having in mind the above mentioned criteria, we decided to base census evaluation on its main objective. That is why special attention was put to essential features of population and housing censuses. According to UN directions, the primary task of the population census is to produce, at regular intervals, the official counting of the population in the territory of a country and in its smallest geographical sub-territories together with information on selected number of demographic and social characteristics of the total population. For this reason, the evaluation provided bellow refers to accuracy in population estimates. The assessment was carried out in terms of demographic analysis of the census data in comparison to previous censuses, administrative data and other existing surveys, especially LFS and mirror statistics.

Evaluation of 2011 Population Census in Poland

There is considerable difficulty in identifying references in assessing the accuracy of the estimates of the population for 2011 census. Population register may serve as one of them. Another may be the census sample survey which was conducted on a random sample of 20% of dwellings on national scale. In the survey, one stage sampling with deep stratification was applied. Out of nearly 13.5 million dwellings, the sample consisted of more than 2 744 thousands. Although for all census results precision tables were provided, the original weights had to be adjusted due to the 13.7% of non-response. The analysis of non-response is not

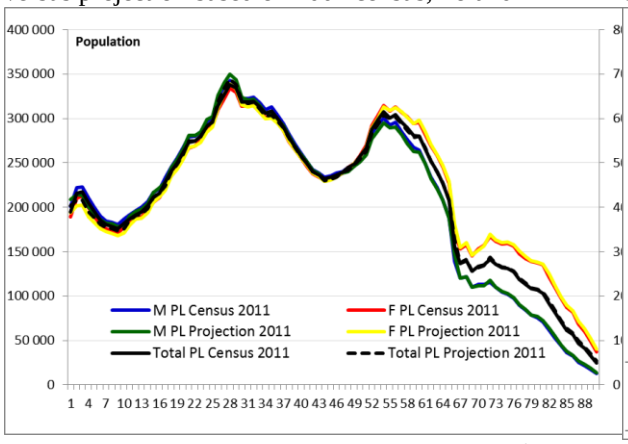
available yet. In turn, census coverage survey which was conducted shortly after the census (from 1 to 11 July 2011), does not meet the requirements of an independent survey carried out in a more precise way. A sample of 80 thousands dwellings was drawn out of 2 744 thousands flats drawn before to the census sample survey. But the frame was restricted only to flats with at least one person with an assigned phone or mobile number. Additionally it covered all dwellings that took part in self enumeration by Internet. Census coverage survey was performed by CATI.

The previous traditional censuses in Poland were evaluated mainly by demographers, who used the possibilities of demographic analysis based on already existing data sources. There is quite well documented evidence on coverage errors in Polish censuses [5][6] [9] [10][12][13]. Among the biggest coverage errors, J. Paradysz [8][9] indicates shortage of up to 30% of women with the shortest length of the marriage (1988 Census), omission of 10% of the youngest infants up to 6 months (2002 Census), omission of the population with increased mobility (2002 Census), lack of elderly (2002 Census).

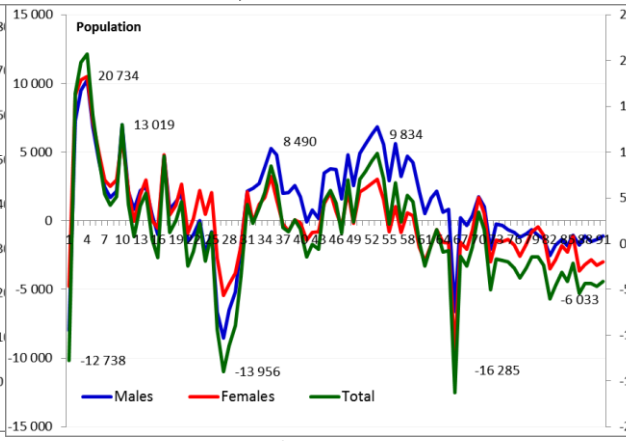
For the purposes of this study, evaluation of last census was based on a comparative analysis of census estimates with other existing sources of information. Because census estimates themselves, as a result of the methodology applied, are obtained with application of comparative analysis, the following thesis was formulated. Due to variety data sources used by mixed-method, census estimates should not be worse than the results based on a single source census. In coverage assessment, special attention was paid to population in age of increased risk of biased estimates. These age groups were defined on the basis of earlier studies [9]. Special consideration was paid to the fact that Poland is a country with strong emigration and consequently to population at age of particularly intensive migration mobility.

The evaluation was performed by using demographic analysis methods. As a starting point population by sex and age of the 2002 census was adopted. Probabilities of survival form life tables for the years 2002-2011 were used to obtain age and sex structure of the population for the following years, similarly as in population predictions. Live births by sex in years 2002-2011 from vital statistics evidence were incorporated and subjected to ageing procedure. Two stage procedure was applied: taking into account infant death (by month) and then life tables probabilities of survival. The procedure did not consider migration, it was applied on national scale and for selected regions.

Graph 1
Population by age and sex, 2011 census estimates versus projection based on 2002 census, Poland



Graph 2
Differences between 2011 census data and projection based on 2002 census, Poland



Source: Polish Population Census 2002 and 2011, life tables and vital statistics (records of births in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia

Comparison analysis of the two population estimates (Graph 1) allows observing good agreement in national scale. The difference between the estimates of the 2011 census and the reference structure is small and amounts about 100 thousands which is 0.26%. However, for women an underestimation is observed and overestimation for men. Different results were also obtained for the different age (Graph 2). These differences allowed identification of age groups requiring special attention: (i) children 0 - 4 years old, (ii) young people: studying and starting their professional career, (iii) working age population, (iv) elderly. The results presented below relate only to: children and adolescents at national scale with some comments from regional analysis. Results obtained for remaining age groups will be presented in further studies.

Table 2
Coverage assessment, differences between 2011 census data and projection based on 2002 census, infants and children

Age	Total		Males		Females		Population aged 0 – 4 years and the difference between estimates		
	persons	%	persons	%	persons	%	2011 Census	Birth Register	Difference
0	-12 846	-3,3	-8075	-4,0	-4 771	-2,5	2057998	1999725	58273
1	16 414	3,8	7 297	3,3	9 117	4,3			2,83%
2	19 776	4,5	9 511	4,3	10 265	4,8			
3	20 734	5,1	10 229	4,9	10 505	5,3			
4	14 086	3,6	6 857	3,4	7 230	3,8			
5	9 192	2,5	4 304	2,3	4 888	2,7			

Source: Polish Population Census 2002 and 2011, life tables and vital statistics (records of births in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia

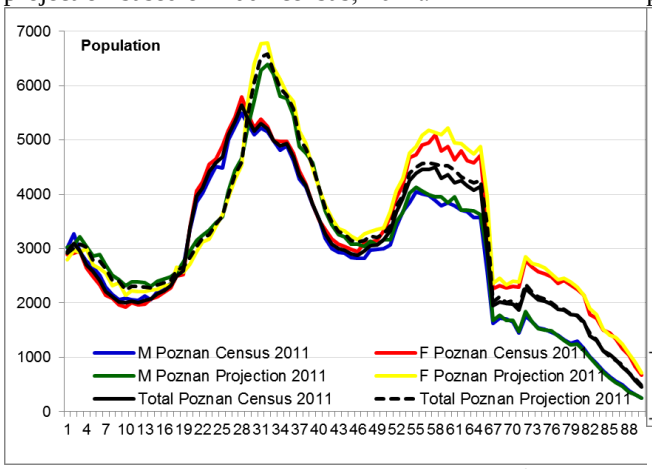
Data obtained for infants show that 2011 census population is underestimated by nearly 13 thousands compared with births and deaths registers (Tab. 2). This represents 3.3% of census number. For children up to 4 years an opposite situation is observed. Census population is overestimated by more than 71 thousands compared with births register. This result is difficult to explain, since census data show children not included in the register. Common mistake is rather underestimating of the population, whereas in this case overall overestimation by 4.2% was observed, and for girls up to 4.6%.

The observed discrepancies might be associated with intensive migration and, discussed by demographers, increased number of births to Polish women abroad, especially in United Kingdom [3][17][19]. Where are the infants registered as born in Poland, but not enumerated during the Census? Where are the children 1-4 enumerated by the census, which were not listed in Polish Birth Register? Answer to these questions is beyond the scope of this paper, but indefinite life situation might suggest Polish migrants to enumerate their children in the census survey in Poland. This might be partly explained [7 p.23] by children born in England and Wales to parents of Polish citizenship acquiring a Polish passport. The ONS data for the youngest age group (0-4 years), show difference amounting almost to 50 thousands between Polish born and Polish nationals. The ONS data show also 74 456 live births in UK to Polish women in 2007-2010 [17 p. 24]. An in-depth mirror statistics might reveal some trends, but it is basically impossible to define the exact numbers.

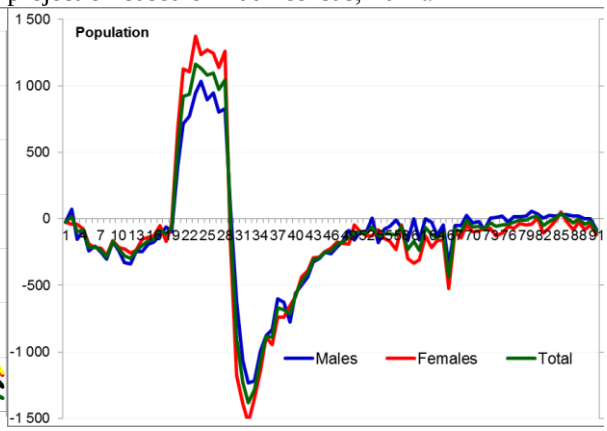
One of the primary purposes of the census is to provide information about age and sex structure of the population in detailed cross-regional division. In case of census based on administrative records, this task is fulfilled, but of course only in respect to information that come from the registers. In relation to these characteristics of the population that can be possessed only from a sample survey, problem of estimating detailed information for small areas arises. This means verification of compliance of definitions and classifications, data integration, estimation for small domains, assessment of consistency of estimates, calibration, etc.

Bearing in mind all the above problems and their consequences, we remain in the analysis of coverage. In regional dimension, coverage assessment is obviously different than for the whole country. In this case, analysis discloses another important problem, which applies to young people, receiving education and starting their professional careers.

Graph 3
Population by age and sex, 2011 census estimates versus projection based on 2002 census, Poznań



Graph 4
Differences between 2011 census data and projection based on 2002 census, Poznań



Source: Polish Population Census 2002 and 2011, life tables and vital statistics (records of births in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia

Exemplary considerations apply to the population of Poznan - the fifth in terms of size city in Poland (554 696). First note greater discrepancies between 2011 census and projections (Graph 3 and 4) at regional than in national scale. The total number of residents of the city is underestimated by more than 20 thousands which is 3.7%.

Table 3
Coverage assessment, differences between 2011 census data and projection based on 2002 census, Poznań

Age	Census		Projection based on 2002 census	'Survival' ratio	Difference between			
	2002	2011			2011 census and adequate 2002 census population		2011 census population and projection based on 2002 census	
					Absolute	Relative (%)	Absolute	Relative (%)
0-4	22 858	22 682	24 923					
5-9	24 827	21 549	23 814					
10-14	31 653	20 652	23 012	0,9035	-2 206	-9,65	-2 360	-11,430
15-19	44 285	26 025	25 804	1,0483	1 198	4,83	221	0,848
20-24	63 232	43 540	33 069	1,3755	11 887	37,55	10 471	24,050
25-29	52 044	53 267	48 949	1,2028	8 982	20,28	4 318	8,106
30-34	36 352	50 637	62 321	0,8008	-12 595	-19,92	-11 684	-23,073
35-39	32 488	41 087	48 145	0,7895	-10 957	-21,05	-7 058	-17,178

Source: Polish Population Census 2002 and 2011, life tables and vital statistics (records of births in the years 2002-2011), http://demografia.stat.gov.pl/bazademografia

The differences between census and projection, are widely disparate according to age (table 3). The biggest underestimate of 11,7 ths. (23,1%) refers to residents of age group 30-34 years. In turn, the greatest overestimation of 10,5 ths.(24,5%) was observed for age 20-24. Similar discrepancies for the age group 20-35 were observed by T. Józefowski and B.

Rynarzewska -Pietrzak [4] who studied quality of population register. As a reason for these discrepancies (amounting to 34%), they pointed relationship between the actual population (census) and permanent residents (register) resulting from Poznan function as a university town which is the capital of the region (students during their studies and taking a job in Poznan after graduation). The current findings seem to be consistent with the analysis conducted for previous census data. Current analysis indicates, however, another problem. It is the decreasing number of city residents, which is not only related to foreign emigration, but also to the process of suburbanization. Within a radius of 20-25 km around the city, new settlements are created and inhabited mostly by young educated people who after graduation and marriage, change student flat in the city into a house near the city. At the moment, this problem is becoming increasingly important.

Conclusion

Great achievement of the census 2011 in Poland was extensive work on evaluation of the quality of administrative records and their use by public statistics. Census based on multiple data sources enforces application of modern methodology. In case of 2011 census in Poland, it meant great scientific work related to the development of modern statistical methods such as calibration, statistical data integration, GIS, estimation for small domains etc. Advantages of the applied methodology are not only difficult to measure and assess, but they should be considered in terms of a precondition for the further development of statistics in the most desirable sense: development of science in response to the needs.

Quality assessment system is built into procedures of conducting a survey that use various data sources. Implementation of mechanism for mutual control, research compliance and comparative analyses results in more reliable information. Use of variety of sources promotes their in-depth exploration also in terms of demographic analyses. On the other hand use of multiple sources of information makes it a ‘natural’ danger of obtaining inconsistent results appears. Divergent estimates, in turn, force attempts to provide consistent estimates and explain reasons of the differences. The 2011 census is a complex procedure combining for each individual, information from two different types of sources: registers which has the character of an inventory, and sample surveys, which examines randomly selected units. This combination implies the need to consider random errors. The analysis using integrated data (register and sample survey) requires development of new theoretical concepts.

Last census in Poland was also an example of the use of modern information and communication technologies in statistical surveys. It allowed reduction of the burden on respondents and enhanced implementation of new channels of data dissemination, website, online database, GIS, web-database. One of the most obvious, for economic reasons, successes of 2011 census in Poland is a significant costs reduction. And last but not least, the most important success of 2011 census, in author's opinion, is a shift in the treatment of census results as an estimate, the quality of which should be assessed and discussed, and not accepted as indisputably certain and true.

Literature

[1] Baldrige M., Brown C.J., Jones S., Keane J.G., 1985, *Evaluating Censuses of Population and Housing*, Department of Commerce, United States of America / US Bureau of the Census

[2] Efron, Bradley, 1978, *Controversies in the foundations of statistics*, "The American Mathematical Monthly" **85** (4): 231–246. doi:10.2307/2321163.

[3] Janta, B. (2013). Polish migrants' reproductive behaviour in the United Kingdom. *Studia migracyjne – Przegląd Polonijny*. 3, 63-96.

[4] Józefowski T. Rynarzewska-Pietrzak B., 2010,*Ocena możliwości wykorzystania rejestru PESEL w spisie ludności,*[w:] *Pomiar i informacja w gospodarce,*Zeszyt Naukowy WIGE, E. Gołata (red.), Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań

[5] Jończy R., 2010, *Migracje zagraniczne z obszarow wiejskich województwa opolskiego po akcesji Polski do Unii Europejskiej. Wybrane aspekty ekonomiczne i demograficzne,* Wydawnictwo Instytut Śląski Sp. z o.o., Opole–Wrocław

[6] Kordos J., 2007, *Some Aspects of Post-Enumeration Surveys in Poland*, „Statistics in Transition – new series", Vol. 8(3), s. 563–576

[7] ONS 2013, *Detailed country of birth and nationality analysis from the 2011 Census of England and Wales*,

[8] Paradysz J., 2002, O błędach nielosowych w badaniu dzietności kobiet w ramach Narodowego Spisu Powszechnego 1970 [w:] *Spisy ludności Rzeczypospolitej Polskiej 1921–2002. Wybór pism demografów*, red. Z. Strzelecki, T. Toczyński, Polskie Towarzystwo Demograficzne, Główny Urząd Statystyczny, Warszawa,s.479–482

[9] Paradysz J., 2010, *Konieczność estymacji pośredniej na użytek spisów powszechnych*,[w:] *Pomiar i informacja w gospodarce*,Zeszyt Naukowy WIGE. E. Gołata (red.), Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu,

[10] Sakson B., 2002, *Wpływ "niewidzialnych" migracji zagranicznych lat osiemdziesiątych na struktury demograficzne Polski*, seria Monografie i Opracowania nr 481, Szkoła Główna Handlowa,Warszawa

[11] Statistics Finland, 2004, *Use of Registers and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland*, Tilastokeskus, Statistikcentralen, Statistics Finland, Helsinki

[12] Śleszyński, P., 2004, *Regionalne różnice pomiędzy liczbą ludności według narodowego spisu powszechnego w 2002 r. i rejestrowaną na podstawie ewidencji bieżącej*. „Studia Demograficzne", 145 (1), s. 93–103

[13] Śleszyński, P., 2005, *Różnice w spisie ludności ujawnione w Narodowym Spisie Powszechnym 2002*, „Przegląd Geograficzny", 77 (2), s.193–212

[14] UN 2006, Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing, 2006, New York, Geneva, United Nations Economic Commissions for Europe, Statistical Office of the European Communities, ECE/CES/STAT/NONE/2006/4, UN Publications,

[15] UN 2012, *Overview of the 2010 round of population and housing censuses in the UNECE region*, ECE/CES/ GE.41/2012/20, 16 May 2012

[16] UN 2013, *Overview Of National Experiences For Population And Housing Censuses Of The 2010 Round*, United Nations Statistics Division, 2013, New York June 2013

[17] Waller, L., Berrington, A. and Raymer, J., 2014, *New insights into the fertility patterns of recent Polish migrants in the United Kingdom*, "Journal of Population Research", DOI: 10.1007/s12546-014-9125-5

[18] Wallgren, A. Wallgren B., 2007, Register-based Statistics. John Wiley & Sons, Ltd.

[19] Zumpe J., Dormon O., Jefferies J., 2012, *Childbearing Among UK Born and Non-UK Born Women Living in the UK*, , ONS, 25 October 2012

[20] Zhang L.-C., 2011, *A Unit-Error Theory for Register-Based Household Statistics*,„Journal of Official Statistics", Vol.27(3), s. 415–432

[21] Zhang L.-C., 2012, Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, vol. 66, no. 1, s. 41–63.

[22] Zhang, L.-C., 2013, Population size estimation based on multiple lists. Uncertainty analysis for categorical data fusion, open lecture given at Poznan University of Economics.

[23] Zynda L., *Lectures on the Philosophy of Science*, that Lyle Zynda taught at Princeton University in the Spring 1994 semester, http://www.soc.iastate.edu/sapp/phil_sci_lecture00.html