This article is accepted for publication in Spatial Economic Analysis

MAIN PAPER

Informal employment in Poland: an empirical spatial analysis

Maciej Beręsewicz[b,c] and D. Nikulin[a]

[a]Department of Economic Sciences, Faculty of Management and Economics, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland; [b]Department of Statistics, Faculty of Informatics and Electronic Economy, Poznań University of Economics and Business, Al. Niepodległości 10, 61-875 Poznań, Poland; [c]Centre for Small Area Estimation, Statistical Office in Poznań, ul. Wojska Polskiego 27/29 60-624 Poznań, Poland

ARTICLE HISTORY
Compiled February 2, 2018

ABSTRACT
The main goal of our article is to bridge the gap in the regional analysis of informal employment in Poland and in particular to indicate the propensity for informal work in the working age population, to test if informal activities are typical for marginalized people (less educated, unemployed, older) and to identify the regional and spatial heterogeneity in the propensity. We use data from the 'Human Capital Balance 2010-2014' survey. Results indicate a strong relationship between the probability of informal work and age, sex and labour force status. Moreover, a strong spatial dependency can be observed.

KEYWORDS
Informal employment propensity; unregistered work; shadow economy; spatial Bayesian analysis; INLA

JEL CLASSIFICATION
J21, J46, R12, R23

1. Introduction

Over the past several decades it has become obvious that informal work is a feature of the contemporary economic landscape, not only in developing countries but also in developed ones. According to Eilat and Zinnes (2002), an understanding of the forces that give rise to the emergence of a shadow economy can help to develop an appropriate policy to deal with it. To answer the question who is affected by informal work, it is necessary to determine the causes of this phenomenon. The core question is why people take informal jobs (Amuedo-Dorantes, 2004). In the literature there is ample empirical evidence indicating which characteristics of workers and employers affect the probability of participating in the informal sector. A comprehensive review of determinants of informal work and the shadow economy can be found in (Kucera & Xenogiani, 2009; Perry, Maloney, Arias, Fajnzylber, & Saavedra-Chanduvi, Jaime Mason, 2007; Schneider & Enste, 2000).

Another issue addressed in this context is whether informal activities should be seen as a way for the poor and marginalized people to cope with poverty, enabling them

CONTACT D. Nikulin. Email: dnikulin@zie.pg.gda.pl

to live a decent life or is it merely a method of tax evasion (Polese & Morris, 2015). In this article, therefore, we aim to determine who is involved in informal activities in Poland and whether people who have a higher propensity for informal work can be really classified as marginalized.

The main motivation for our research is the scarcity of studies on informal employment at a more disaggregated level (sub-national level). Obviously, the main obstacle is data availability. However, it is argued that policies aimed at reducing informal employment should be adapted to different regions (Herwartz, Tafenau, & Schneider, 2015). Existing studies on regional differences in the scale of the shadow economy and informal employment indicate a considerable variation in the extent of shadow activities within countries (Almeida & Carneiro, 2009; Di Caro & Nicotra, 2016; Herwartz et al., 2015; Jonasson, 2012; Tafenau, Herwartz, & Schneider, 2010). To the best of our knowledge, there is no study on regional disparities of informal employment in Poland.

Using regional data, we identify spatial correlations of patterns of the shadow labour market in Poland. In particular, the objectives of this study are (1) to investigate who is affected by informal work, (2) to test if informal activities are typical for marginalized people (less educated, unemployed, older), (3) to identify regional and spatial patterns in the extent of informal employment in Poland. To do this, we use data from the *Human Capital Balance* survey (hereafter the BKL survey) from the years 2010-2014, which covers the working age population. To uncover spatial associations, we apply Bayesian hierarchical models using the Integrated Nested Lagrange (INLA) approach proposed by Rue, Martino, and Chopin (2009). Our analysis shows that there is a strong relationship between the probability of informal work and age, sex and labour force status. Moreover, a strong spatial dependency can be observed.

The rest of this article is organised into five sections. In the second section we provide a thorough review of literature on informal employment and regional studies on the labour market. The third section focuses on the data, including a description of the study and initial spatial analysis. In the fourth section we describe the models used for regression modelling and results of our analysis. The article ends with conclusions and possible directions for future research.

2. Relevant literature

2.1. Informal employment in the light of the marginalization thesis

The literature does not conclusively settle the question whether informal employment should be seen as a way for the poor and marginalized people to cope with poverty (Polese & Morris, 2015). According to one view, informal jobs are jobs with less favourable working conditions. This structuralist standpoint assumes that informal work involves low-paid, insecure, unregulated jobs carried out by marginalized people (Ahmad, 2008; Gallin, 2001). The view that undeclared employment is closely related to poverty and mainly affects marginalized people is what some studies refer to as the 'marginalization thesis'. Numerous studies on this topic conducted by Williams and Horodnic (2015a, 2015b, 2015c) explain which part of the population in European countries is predominantly engaged in the informal sector. Analysing the group of self-employed they find that the marginalization thesis is valid when one considers such characteristics as age, marital status, attitudes towards tax compliance, occupation and financial circumstances of households. At the same time factors such as the urban-rural divide and the educational level do not influence the propensity for informal

work (Williams & Horodnic, 2015c). Their other studies confirm the marginalization thesis to some extent. The authors show that within a given population some groups of people have a higher propensity to work informally, e.g. the unemployed, people who struggle to pay their household bills, younger age groups and women (Williams & Horodnic, 2015b), people defining themselves as working class and those who hold non-conformist beliefs about tax compliance (Williams & Horodnic, 2015a). However, based on different data sources, the results of the above studies are ambiguous and do not provide clear conclusions.

2.2. Informal employment in Poland

In this section we present previous research on informal employment in Poland in order to show its determinants. Cichocki and Tyrowicz (2010) examine disparities between wages of informal and formal workers in order to find reasons for informal employment. Based on survey data from the Centre for Social and Economic Research and the Polish Ministry of Labour and Social Affairs, Cichocki and Tyrowicz (2010) find that incomes of informal workers are lower than those earned by workers in formal employment. Thus, the authors confirm to some extent the labour market segmentation hypothesis, which holds that some workers are absent from the official labour market owing to access costs or demand constraints. However, Tyrowicz and Cichocki (2011) using their own definition of unregistered employees (people registered as unemployed but holding an informal job) and employing data from the Polish Labour Force Survey (LFS), compare higher wages received from undeclared work with those earned in formal employment. Their findings provide support for the tax evasion hypothesis rather than the dual labour market hypothesis. It should be noted, however, that the above two studies have different reference periods and concern different groups of informal workers. In fact, they both acknowledge the heterogeneity of informal workers and the need for further investigations. To explain the reasons for informal work the authors also examine the propensity for undeclared employment (Cichocki & Tyrowicz, 2011). They find that workers with lower qualifications have a greater propensity for informal work. Moreover, micro firms and companies in the construction, agriculture and trade industries are more likely to opt for informal employment. In contrast, such determinants as age or rural location are not significant in explaining the propensity for informal work. In addition, characteristics of informal workers vary throughout the business cycle. During a period of economic prosperity informal work is mostly chosen by less skilled people and those in traditional sectors; when the economic situation deteriorates, more people, including those with better qualifications, are likely to work informally.

Another source of data on informal employment in Poland is the Eurobarometer survey, which was conducted by the European Commission (European Commission, 2007, 2014) in 2007 and in 2013. Results of these studies can be divided into two parts: the first one focuses on explaining the cross-country disparities in informal work. Williams (2015a) analyses Central and East Europe (CEE) countries and indicates that in more developed, less corrupt and more equal economies with a higher level of taxation, social protection and more effective redistribution via social transfers, envelope wages are less popular and are mainly associated with overtime work. Moreover, analysing the prevalence of informal employment Williams (2015b) shows that in wealthier, less corrupt and more equal economies with higher level of taxation, social protection and more effective redistribution via social transfers, the level of informal employment is

also lower. The second type of analysis based on the Eurobarometer survey is devoted to explaining the nature of informal employment. The study carried out by Williams and Horodnic (2015b) indicates that some marginalized people like younger workers, those with fewer years of formal education, unemployed, single-person households, people struggling to pay bills are more likely to work informally. On the other hand, they find no association between a higher propensity to work informally and marital status, social class, the number of children or residence in rural areas.

The study conducted by Łapiński, Peterlik, and Wyżnikiewicz (2015) shows that informal employment is mainly performed by marginalized people, such as unemployed with limited opportunities to find work because of poor qualifications, pensioners, students and pupils, immigrants and generally people with low qualifications. A study performed by the Central Statistical Office in Poland (Central Statistical Office, 2015) indicates that informal work tends to be performed by men rather than women, people with a lower educational level, aged 35-44 and 45-59 but is equally distributed between inhabitants of villages and towns.

2.3. Regional analysis

With regard to high variation in labour market outcomes within one country, it seems necessary to conduct analysis at a more disaggregated level. In the current literature there is a lot of research confirming the view that the regional perspective is becoming more and more important (Fischer & Nijkamp, 2014, p. xxi). Spatial relations play an important role in creating labour market outcomes (Fernandez & Su, 2004; Kelly, 2011). Coe, Kelly, and Yeung (2013, p. 161) argue that labour markets are determined by social institutions in a given region. The reduction of barriers in international trade and the free transfer of materials, technologies, people and capital all contribute to a higher mobility between regions (Coe et al., 2013, p. 156-157). The literature review provides a long list of spatial aspects affecting labour market outcomes, like the regional diversity in employment and unemployment levels (Ciżkowicz, Kowalczuk, & Rzońca, 2016; Marelli, Patuelli, & Signorelli, 2012; Newell & Pastore, 2006; Novotný & Nosek, 2012; Patuelli, Schanne, Griffith, & Nijkamp, 2012). Following Fischer and Nijkamp (2014, p. xxviii), who claim that 'spatial interdependencies have always been at the heart of regional science research', we decided to focus not only on individual factors influencing the probability of work informally but also to consider spatial effects. So far regional analyses of informal employment are rather scarce and focus mostly on local disparities in the extent of the phenomenon. Jonasson (2012), using Brazilian worker-level data (Brazilian Demographic Census for the year 2000), find that the probability of a worker being employed informally is lower in regions with better governance and higher average education. In this way the author underlines the role of government effectiveness in creating conditions for informal activities and confirms regional determinants affecting the individual employment outcomes (Jonasson, 2012). At a sub-national level, there are a few empirical studies concerned with regional variation in the size of the informal sector, for example, Chaudhuri, Schneider, and Chattopadhyay (2006) analyse 14 states in India and Torgler and Schneider (2007) examine 26 cantons of Switzerland. Herrera-Idárraga, López-Bazo, and Motellón (2016) using Colombian Household Survey find that the extent of informality may influence the regional wage gap differentials especially at the bottom part of wage distribution. At the micro level, few empirical studies analyse the effect of institutional factors on the propensity of workers or businesses to participate in the informal sector. A recent

exception is the study by Almeida and Carneiro (2009), who analyse how differences in enforcing labour regulation across regions in Brazil affect regional informal employment and unemployment.

Considering the above, we attempt to include the impact of spatial relations on the level of informal employment in Poland. As pointed out by Fernandez and Su (2004), there is still a need for a description of spatial factors influencing labour market outcomes. With this in mind, we aim to answer the question of whether informal work is really the last resort for marginalized people. Moreover, we want to consider marginalization from the individual and spatial point of view in order to examine if people from less developed regions in Poland are more willing to work informally. To the best of our knowledge it is the first study to combine the explanation of individual socio-economic factors influencing the propensity for informal work with the impact of spatial effects.

3. The data

3.1. Description of the study

As the scope of the concept of informal employment is very wide, it is necessary to establish theoretical boundaries of this phenomenon. Many institutions address this issue in order to analyse its character and to formulate appropriate policy measures. In this paper we use the definition proposed by OECD, according to which informal employment covers (Venn, 2008):

- employees unregistered for mandatory social security,
- employees paid less than the minimum wage,
- employees without a written contract (if it is required),
- employees and self-employed who hide or understate their income,
- unregistered firms and their employees,
- "false self-employed".

Taking into account the definition proposed by OECD and the available data set, in this study informal employment is described as *employment without a written contract.* We use data from the BKL survey for the years 2010-2014. The choice of the reference period is motivated by data availability. In the article we focus on the working age population, women aged 18-59 and men aged 18-64 and the NUTS 3 level (Nomenclature of Territorial Units for Statistics, 66 subregions). Selection of this level is motivated by: (1) insufficient sample size at LAU 1 (Local Administrative Units), (2) insufficient number of representatives of informal workers and (3) the fact that not all LAU 1 units are observed in the reference period. Therefore, we have decided to conduct our analysis at a higher level of aggregation. A detailed description of the study is presented in Section A in online supplementary material.

Table 1 presents Horvitz and Thompson (1952) direct estimates of the share of all categories of participation in the labour market in Poland between 2010 and 2014 based on the BKL survey. The percentage is stable over time, close to 4% and consistent with Central Statistical Office (CSO) reports on unregistered work, which was equal to 4.6% in 2010 and 4.5% in 2014 Central Statistical Office (2015). However, the comparison with CSO data should be taken with caution in view of the different definitions of the statistical population (CSO analysed the population of persons aged 15+).

[Table 1 near here].

Table 2 presents the descriptive statistics of direct estimates of the share of informal workers at NUTS 3 level between 2010 and 2014. Names, numbers and locations of each subregion are presented in Figure F1 in online supplementary material. The distribution is slightly skewed to the right because of the maximum values. The descriptive statistics for each year are similar except for 2010, when the maximum value was equal to 17% (for Ełcki subregion, no. 55), which is certainly an outlier. The same problem involves three other subregions (Leszczyński, no. 59 in 2010, Rybnicki no. 49 in 2011 and 2012, Pilski no. 60 in 2012), none of the respondents in the sample provided a positive answer to the question about an informal agreement. This can be attributed either to sampling error or to non-response, but it cannot be determined based on available BKL data and reports. There are subregions in which informal workers accounts for close to 10% of the working age population. In each year we observe variability between subregions, which will be further investigated in terms of spatial effects.

[Table 2 near here].

Based on the literature of the subject, we select the following variables for modelling: (1) respondent's sex (0 = Male, 1 = Female), labelled `sex`, (2) age categorized into 5 groups (18-24, 25-34, 35-44, 45-54, 55-59 for females and 55-64 for males), labelled `age5`, (3) education categorized into 4 levels (1 = Primary, 2 = Basic vocational, 3 = Secondary, 4 = Tertiary), labelled `educ4`, (4) labour force status according to ILO (1 = Unemployed, 2 = Working, 3 = Inactive), labelled `LFS_status`, (5) whether a given person had children (1 = Yes, 0 = No) labelled `children`, (6) area of residence categorized into 6 classes (1 = Rural area, 2 = Urban area to 50k, 3 = Urban area between 50k and 100k, 4 = Urban area between 100k and 200k, 5 = Urban area between 200k and 500k and 6 = Urban area over 500k), labelled `locality` and (7) part-time work (1 = Yes, 0 = No) labelled `part_time`. Moreover, we add one interaction between age and sex, which is motivated by the previous literature (Tyrowicz & Cichocki, 2011). We also include one covariate at NUTS 3 level which represents the ratio of the number of long-term registered unemployed (over 12 months) to all registered unemployed, which is labelled `register_unempl >= 12m`. This variable is centered at its overall mean. Finally, we also include information about the trend in the model (1 = 2010, ..., 5 = 2014), labelled `trend`.

[Figure 1 near here].

Figure 1 and Table A2 (in online supplementary material) present exploratory data analysis of the demographic variables selected for the modelling procedure based on pooled samples. The highest proportion of informal workers can be observed in the group of part-time workers (around 12.4%), unemployed persons (around 10.5%), people aged 18-25 (around 7.5%), those with only primary education (around 7%) and those without children (round 6%). In terms of the area of residence, the share of informal workers in rural and smallest urban areas is lower than the overall average, whereas it is higher for the biggest urban areas.

3.2. Initial spatial analysis

Figure 2 presents initial spatial analysis of the overall share of informal workers between 2010 and 2014. According to the BKL survey, the highest share of informal workers

is recorded in the north-eastern part of Poland, in particular in Suwalski subregion (no. 39, 8.9%), Białostocki subregion (no. 37, 8.0%) and Łomżyński subregion (no. 38, 7.4%). We can roughly divide the country into two parts: one including eastern, central, and south-eastern part with a relatively higher level of people engaged in informal work, and the north-western part, with a lower proportion of informal workers. This division is very approximate and there are several exceptions. Interestingly, in some bigger cities, such as Wrocław (no. 5), Kraków (no. 21), Szczecin (no. 65) and Łódź (no. 16), the share of informal workers is relatively high (accordingly 5.63%, 7.29%, 5.47%, 4.79%). In the south of Poland there are also subregions with a high level of informal employment, such as Katowicki (no. 48, 6.9%), Krakowski (no. 20, 5.38%), Sosnowiecki (no. 50, 5.26%), Tarnobrzeski (no. 36, 4.77%). Subregions with the lowest share of unregistered workers include Leszczyński (no. 59, 1.1%), Rybnicki (no. 59, 1.2%) and Bielski (no. 44, 2.0%). However, it should be noted that in the first two subregions some annual samples do not include a single case of informal work.

[Figure 2 near here].

In general, based on the initial spatial analysis of informal workers in Poland, it is hard to recognise clear regional patterns although there are clusters of subregions with a high share of informal workers. That is why a more in-depth analysis of reasons for undeclared work is required. At this stage, we cannot conclude that the level of informal employment is purely related to the location of a given region.

However, as the data are of hierarchical structure we should account for dependence and variance heterogeneity within and between podregions. In general, omission of this characteristic may lead to inefficiency or biased parameter estimates. Moreover, a structure of the covariance should be taken into account as wrongly selected might lead to model misspecification. For instance, the omission of the spatial component in the model is likely to render the estimations inefficient and inaccurate in comparison to unstructured random effect. This fact should not be underestimated, since, after all, the presence of spatial effects violates the assumption of independently and identically distributed (i.i.d.) errors of most statistical procedures, and they can even invert the slope of estimated coefficients from non-spatial analysis, which may inevitably lead to false and wrong conclusions. In the next sections we will therefore analyse factors which increase the probability of informal work based on individual data and verify existence of spatial effect.

4. Regression modelling

4.1. Models discussed in the article

The modelling procedure is conducted as follows. First, we poll data from all years into one dataset. Then, we study the propensity to work informally, which is estimated based on the target variable assumed to follow Bernoulli distribution and defined as in equation (1)

$$y_{it} = \begin{cases} 1, & \text{if respondent declared working without a formal agreement,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $i = 1, ..., m$ denotes respondent identifier and $t = 1, .., 5$ denotes wave of the BKL

survey. Let ρ be the propensity to work informally that is the conditional probability that an individual declares to work without an formal agreement ($y = 1$) given observed characteristics $\boldsymbol{x}_{id}$. We treat model given by equation (2) as a starting point for our process process

$$\rho(\boldsymbol{x}_{it}) = Pr(y_{it} = 1|\boldsymbol{x}_{it}). \tag{2}$$

To estimate (2) we use a logit link function $g()$, namely

$$g(\mu_{it}) = \text{logit}(\mu_{it}) = \log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \eta_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta}, \tag{3}$$

where $\mu_{it} = E(y_{it}|\boldsymbol{x}_{it}) = Pr(y_{it}|\boldsymbol{x}_{it})$. This is a generalized linear model (GLM).

Prior the selecting the model, we verify the ignorability of the sampling scheme applied in the BKL survey, which is described in detail in Section B in online supplementary material.

To investigate the spatial effect, we apply the following models: a generalized linear model with i.i.d random effect (GLMM), GLMM with spatially correlated random effects (BESAG; Besag 1972) and GLMM with i.i.d and spatially correlated random effect (BYM; Besag, York, and Mollié 1991). To estimate these models we use Integrated Nested Lagrange (INLA) introduced by Rue et al. (2009). This approach is implemented in `R-INLA` package (Lindgren & Rue, 2015).

The GLMM model extends the (3) model by adding an unstructured random effect denoted by $\boldsymbol{u}_d$ associated with NTS3 level denoted by $d = 1, ..., D$ and $D = 66$. Random effect u_d is assumed to be normally distributed. This model is given by equation (4):

$$\begin{aligned} g(\mu_{it}) &= \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{u}_d, \\ \boldsymbol{u}_d &\sim \mathcal{N}(\boldsymbol{0}, \tau_u^{-1}\boldsymbol{I}), \end{aligned} \tag{4}$$

where τ_u is the conditional precision of the random effect. By default R-INLA assigns a log-gamma distribution prior to $\log(\tau_u)$.

Further, we consider two models (BESAG and BYM) that include the intrinsic conditional autoregressive (CAR) specification. BESAG assumes that the random effect has a CAR specification, that replaces $\boldsymbol{u}_d$ from equation (4) with $\boldsymbol{v}_d$ given by (5)

$$\begin{aligned} g(\mu_{it}) &= \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{v}_d, \\ \boldsymbol{v}_d &\sim \mathcal{N}(\boldsymbol{0}, \tau_v^{-1}\boldsymbol{Q}^-), \end{aligned} \tag{5}$$

where $\boldsymbol{Q}^-$ denotes the generalized inverse of $\boldsymbol{Q}$ and $\boldsymbol{Q}$ is the precision matrix related to neighbour structure that is defined in equation (6)

$$Q_{dl} = \begin{cases} \eta_{\delta d} & d = l, \\ -1 & d \sim l, \\ 0 & \text{else}, \end{cases} \tag{6}$$

where δd denotes the set of neighbours of region d, $\eta_{\delta d}$ denote its size and $d \sim l$ denotes that area d and l are neighbours if they share a common border (Riebler, Sørbye, Simpson, & Rue, 2016). The BYM model contains two random effects and is defined as follows

$$\begin{aligned} g(\mu_{it}) &= \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{u}_d + \boldsymbol{v}_d, \\ \boldsymbol{u}_d &\sim \mathcal{N}(\boldsymbol{0}, \tau_u^{-1}\boldsymbol{I}), \\ \boldsymbol{v}_d &\sim \mathcal{N}(\boldsymbol{0}, \tau_v^{-1}\boldsymbol{Q}^-), \end{aligned} \tag{7}$$

where symbols are defined as previously. To estimate random effects we need to establish the joint distribution π, which only for $\boldsymbol{v}_d$ is given by equation (8)

$$\pi(\boldsymbol{v}_d|\tau_v) \propto \exp\left(-\frac{\tau_v}{2}\boldsymbol{v}_d^T\boldsymbol{Q}\boldsymbol{v}_d\right). \tag{8}$$

For detailed description of estimation of parameters in INLA please refer to Rue et al. (2009).

Further, in order to select the most suitable model we use the deviance information criterion. It is the most commonly used measure of model quality and is based on the deviance measure and the number of effective parameters

$$DIC = \bar{D} + p_D, \tag{9}$$

where $\bar{D}$ is is the mean of the Bayesian deviance and p_D is an effective number of parameters which is proportional to the deviance variance and is regarded as a measure of model complexity. Similarly to AIC and BIC, models with smaller DIC are better supported by the data. Finally, we calculate Watanabe-Akaike information criterion (WAIC). WAIC is a more fully Bayesian approach for estimating the out-of-sample expectation starting with the computed log point-wise posterior predictive density and then adding a correction for the effective number of parameters to adjust for overfitting (Gelman, Hwang, & Vehtari, 2014).

4.2. Results of the analysis

Table 3 contains information criteria and estimated hyperparameters (variance components) for the models described in the previous section. The latter are described in terms of the median and 95% credible intervals. DIC and WAIC suggest that GLMM describes the decision to take up informal work better in comparison with GLM. Further, information criteria suggest that the models with the spatial random effect are better compared to the model with the i.i.d random effect. In the case of the BYM model, the variance of the spatial effect dominates the unstructured random effect. In both spatial models, the variance of the random effect is higher compared to the mixed effects model with i.i.d. However, information criteria indicate that BYM model should be used. Further, we compare estimates of fixed effects and random effects in order to verify whether omission of random effects leads to inefficiency or biased parameters. See Figure C1 and D1 in online supplementary material. Results for fixed effects indicate that slight bias is observed only for `trend` and big cities (`locality` variable),

and for only for `locality` credible intervals are longer for BESAG and BYM models. For random effects differences in point estimates are visible in particular for the podregions with the highest propensity to work informally however credible intervals overlap. Finally, the distribution of the propensity for informal work among people of working age is presented in Section E in online supplementary material.

[Table 3 near here].

[Table 4 near here].

Based on the BYM model (results from Table 4), we can assess the probability of taking up informal work with regard to a rich array of socio-economic factors. With regard to one of the objectives of our study, namely the verification of the marginalisation thesis, we verify if people who can be described as marginalized really have a greater propensity for informal work.

According to the estimated BYM model, we find that people who are unemployed show a significantly higher propensity for informal work than those who work. Our results are therefore consistent with previous studies on informal employment in CEE countries (Williams & Horodnic, 2015b). Moreover, those who are inactive in the labour market are more likely to take up an informal job than employed people. This could mean that informal jobs are more prevalent among people with fewer opportunities in the labour market. To some extent, this conclusion is consistent with the statement about informal employment as last resort jobs (Harris & Todaro, 1970).

Further evidence in support of the above hypothesis is that part-time workers are more likely to work informally. We can therefore assume that people who are somehow marginalized in the labour market generally show a higher propensity for informal work. Importantly, marginalization in the labour market is obviously related to human capital theory. In this field our findings show that people with the primary level of education are more likely to work informally than people with basic vocational, secondary and tertiary education. The higher the educational level, the smaller the likelihood of working informally. Considering other social characteristics of people involved in informal activities, we can see that men have a higher propensity for informal work than women. Similar results are obtained in a study conducted by Williams and Horodnic (2015b). Moreover, analysing sex-age interactions, the outcome is that younger people (18-24 years) have a higher propensity for informal work than older age groups. In other words, the probability of informal work decreases with age. The interactions also confirm the previous evidence about a higher propensity for informal work among men than among women. This relationship occurs in every age cohort.

On the other hand, the variable 'locality' indicates that people living in urban areas are more likely to work informally than people from rural areas, which is consistent with findings presented by Williams and Horodnic (2015b). In particular, we find that residents of medium-sized towns with a population of 50-200 thousand inhabitants are most likely to work informally. Furthermore, people who declare having children have a lower propensity for informal work compared to those without children. This could be related to the fact that informal employment is most prevalent among young people, who tend not to have children. Since our data cover the period of 2010-2014 we are able to observe changes over time. Between 2010 and 2014 there is a decreasing trend in propensity for informal work. Based on regional data, we can see how the situation in the local labour market (measured by the rate of log term registered unemployed to all registered unemployed) influences individual propensity for informal work. We find therefore that an increase in the share of long term unemployed in subregions

is associated with a slightly higher probability that people from these subregions will be working informally. This means that a less favourable situation in the local labour market due to long-term unemployment slightly increases the propensity for informal work. With respect to the statement that unemployed people are more likely to take up an informal job (what results from our model), it can be concluded that informal work is related to unemployment. In general, a person's decision to take a job in the informal sector is determined by their individual characteristics, current labour status and the external situation in the local labour market.

[Figure 3 near here].

In the next step, we analyse spatial patterns in the propensity for informal work across the country. Figure 3 presents the distribution of the spatial random effect estimated from the BYM model and the probability that the level of the random effect is greater than zero ($Pr(v_d > 0)$, where v_d is a spatial random effect estimated in the BYM model). Positive values of this effect indicate high propensity for informal work, while values below zero indicate that persons from these regions are less likely to work informally. Figure 3 shows spatial heterogeneity in the propensity for informal work.

We create five intervals to describe the probability of the positive effect on propensity with the following number of cases: 23 subregions in $(0, 0.2]$, 10 subregions in $(0.2, 0.5]$, 13 subregions $(0.5, 0.8]$, 11 subregions in $(0.8, 0.95]$ and 9 subregions $(0.95, 1]$. The results are presented in the bottom map in Figure 3. In general, a positive propensity for informal work is observed in the eastern part of Poland, particularly in the north-eastern subregions. There is also another cluster in the south, namely the city of Krakow and its surrounding districts.

The results indicate the existence of significant spatial patterns in the propensity for informal work. We find that besides individual characteristics which increase the likelihood of informal employment, the spatial aspect also plays an important role. The place of residence of a given individual has an impact on their willingness to engage in informal work. However, looking at the regional differences in the probability of informal work, it is difficult to give a straightforward answer. In order to investigate possible explanations, we analyse several economic variables which characterize the regions. All materials can be found in the online supplementary material. Section G presents the current regional situation at NUTS 3 level and covers relevant labour market variables as well as GDP per capita. Looking at the maps, regional differences in the share of informal workers could be explained as follows. First of all, the greater propensity to work informally may result from the relation between the share of self-employed and paid employees in total employment. Figure G1 shows that the regions along the eastern border of Poland are characterized by a greater prevalence of self-employment and, simultaneously, by a lower share of paid-employees. It can therefore be argued that regions with a higher share of self-employment and a lower share of paid employees can be identified as those with a greater prevalence of informal workers. This explanation assumes that the phenomenon of self employment is often related to informal work since people from areas with fewer opportunities in the legal labour market may be forced to be active as self-employed (often bogus self-employment status) as well as work informally. Partially, this explanation is related to the assumption that informal work is more prevalent in poorer regions. To prove this we illustrate the spatial structure of Gross Domestic Product per capita (GDP per capita), which is presented in Figure G3. We can observe a rather negative correlation between the level of development of a given region (expressed through GDP per capita) and the share of informal

workers. Similar results are obtained by Nikulin and Sobiechowska-Ziegert (in press) who analyse the determinants of the prevalence of informal workers in Polish provinces (NUTS 2). Moreover, the spatial distribution of unemployment (Figure G2) indicates a positive relation between unemployment and the scope of informal work. In particular, there is a significant correlation between the share of long-term unemployed (upon 12 months) and the share of informal workers (correlation coefficient = 0.0743). When the impact of the unemployment rate is analysed, the relation is no longer so clear. This could indicate that in regions where long-term unemployment is more prevalent (and thus the share of less skilled people is larger) the probability of informal work is greater.

5. Conclusions and further research

The literature review provides a wide range of reasons for and determinants of informal employment. Because of the complexity of this phenomenon, so far no unambiguous explanations have been suggested. Moreover, many empirical studies deliver different results, which makes it difficult to draw straightforward conclusions. On the other hand, policy makers want to know determinants of informal employment in order to develop comprehensive policy measures. For these reason, there is still a need for more in-depth studies of informal employment, especially regarding its nature and characteristics.

The purpose of our analysis is twofold. Firstly, we aim to analyse people's propensity for informal work, with an emphasis on the marginalized part of society. Secondly, we use sub-regional data in order to indicate spatial patterns in the level of engagement in informal work. Using individual data we could describe the propensity for informal work depending on socio-economic characteristics and location. We find that unemployed, economically inactive, younger people, with a lower educational level, or engaged in part-time work are more likely to take up an informal job. In this way we confirm the marginalization thesis to some extent, since the above mentioned groups are mostly perceived as marginalized in society and in the labour market in particular. Moreover, men are more likely to work informally than women, in each age cohort. Interestingly, people from medium-sized towns have a higher propensity for informal work, which is at the lowest in rural areas. Taking into account the random spatial effect, we find that people living in the north-east of Poland have a higher propensity to take an informal job. Moreover, sub-regions with a higher probability of informal work tend to cluster. The regional disparities in propensity to informal work may be partially explained by the differences in regional structure in self-employment share as well as the level of development of given region (approximated by the GDP per capita). We find that in less developed subregions, where the self-employment is more prevalent and the share of long term unemployed is greater the probability of taking informal job is higher.

Our results may indicate important advice towards the policy issues in several ways. First of all, we provide an empirical evidence that informal work is often related to marginalization in the society and in the labour market in particular. Since for young and poor educated people the opportunities in the legal labour market are limited, they may be more inclined to look for alternatives in the informal market. The policy makers should be aware of the groups in the society which are most likely to engage in informal sector. A greater attention should be paid to those marginalized people in order to recognize their constrains and suit proper measures i.a. through enhancing the education or qualification level. Secondly, our results show that unemployed people are more likely to work informally. Therefore, a proper policy reducing unemployment

may bring significant results in the decrease of informal work. Thirdly, we point out that informal activities are prevalent not only in small towns and in rural areas, what may be expected based on labour marker opportunities. A significant part of informal work takes place in medium size town and thus it is important to not overlook or underestimate this phenomenon in those areas while directing the policy measures.

However, given that our study is based on survey data, it has some limitations that should be taken into consideration. The first problem is non-response, which may be connected with the research topic. However, because of the lack of auxiliary data sources, it is impossible to verify whether the underlying mechanism of non-response is random or non-random (Rubin, 1976). The second problem is the measurement error related to both the target variable (false responses) and auxiliary variables (labour status, education) that were not verified by interviewers. Only access to unit-level data that can be linked to official data sources can provide information about the level of errors in these variables (except the target variable). Finally, the propensity for informal work is skewed, which calls for further analysis using robust methods or M-quantile based approach (cf. Chambers, Chandra, Salvati, & Tzavidis, 2014).

In summary, our study fills a research gap concerning the analysis of individual determinants of informal work based on sub-regional data. Our main contribution is the inclusion of spatial effects to account for the propensity to work informally. Our results are a good source of knowledge about people engaged in informal work, which can be used to inform programs aimed at combating or limiting the scope of the undeclared economy.

Acknowledgement(s)

The authors are grateful to the Editor-in-Chief and the reviewers for careful reading of a former draft of this paper and providing constructive comments, which significantly improved the content of this paper.

References

Ahmad, A. N. (2008). Dead men working: time and space in London's ('illegal') migrant economy. *Work, Employment and Society*, *22*(2), 301–318. doi: 10.1177/0950017008089106.

Almeida, R., & Carneiro, P. (2009). Enforcement of labor regulation and informality. *Journal of Comparative Economics*, *37*(1), 28–46. doi: 10.1016/j.jce.2008.09.004.

Amuedo-Dorantes, C. (2004). Determinants and Poverty Implications of Informal Sector Work in Chile. *Economic Development & Cultural Change*, *52*(2), 347–368. doi: 10.1086/380926.

Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 75–83. Retrieved from `http://www.jstor.org/stable/2985051`

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20. doi: 10.1007/BF00116466.

Central Statistical Office. (2015). *Praca nierejestrowana w Polsce w 2014 roku [Unregistered Employment in Poland in 2014].* Warsaw: Central Statistical Office. Retrieved from `https://stat.gov.pl/obszary-tematyczne/rynek-pracy/opracowania/praca-nierejestrowana-w-polsce-w-2014-roku,4,4.html`

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

76(1), 47–69. doi: 10.1111/rssb.12019.

Chaudhuri, K., Schneider, F., & Chattopadhyay, S. (2006). The size and development of the shadow economy: An empirical investigation from states of India. *Journal of Development Economics*, *80*(2), 428–443. doi: 10.1016/j.jdeveco.2005.02.011.

Cichocki, S., & Tyrowicz, J. (2010). Shadow employment in post-transition-Is informal employment a matter of choice or no choice in Poland? *Journal of Socio-Economics*, *39*(4), 527–535. doi: 10.1016/j.socec.2010.03.003.

Cichocki, S., & Tyrowicz, J. (2011). Determinanty zatrudnienia nierejestrowanego w Polsce w okresach wysokiej i niskiej koniunktury gospodarczej [Determinants of undeclared employment in Poland in periods of high and low economic prosperity]. *Gospodarka Narodowa*, *3(235)*. Retrieved from `http://gospodarkanarodowa.sgh.waw.pl/p/gospodarka{_}narodowa{_}2011{_}03{_}01.pdf`

Ciżkowicz, P., Kowalczuk, M., & Rzońca, A. (2016). Heterogeneous determinants of local unemployment in Poland. *Post-Communist Economies*, *28*(4), 487–519. Retrieved from `http://dx.doi.org/10.1080/14631377.2016.1226784`

Coe, N. M., Kelly, P. F., & Yeung, H. W. (2013). *Economic geography. An contemporary introduction.* Hoboken: NJ: John Wiley & Sons.

Di Caro, P., & Nicotra, G. (2016). Short, Long and Spatial Dynamics of Informal Employment. *Regional Studies*, *50*(11), 1804–1818. doi: 10.1080/00343404.2015.1072274.

Eilat, Y., & Zinnes, C. (2002). The shadow economy in transition countries: Friend or foe? a policy perspective. *World Development*, *30*(7), 1233–1254. doi: 10.1016/S0305-750X(02)00036-0.

European Commission. (2007). *Undeclared Work in the European Union. Special Eurobarometer 284* (No. October). Brussels: European Commission. Retrieved from `http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs{_}284{_}en.pdf`

European Commission. (2014). *Undeclared work in the European Union. Special Eurobarometer 402.* Brussels: European Commission. doi: 10.2767/37041.

Fernandez, R. M., & Su, C. (2004). Space in the Study of Labor Markets. *Annual Review of Sociology*, *30*, 545–569. doi: 10.1146/annurev.soc.29.010202.100034.

Fischer, M. M., & Nijkamp, P. (2014). *Handbook of Regional Science.* London: Springer Reference.

Gallin, D. (2001). Propositions on Trade Unions and Informal Employment in Times of Globalisation. *Antipode*, *33*(3), 531–549. doi: 10.1111/1467-8330.00197.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi: 10.1007/s11222-013-9416-2.

Harris, J. R., & Todaro, M. P. (1970). Migration, unemployment and development: a two-sector analysis. *The American Economic Review*, *60*(1), 126–142. Retrieved from `http://www.jstor.org/stable/1807860`

Herrera-Idárraga, P., López-Bazo, E., & Motellón, E. (2016). Regional Wage Gaps , Education and Informality in an Emerging Country: The Case of Colombia. *Spatial Economic Analysis*, *11*(4), 432–456. doi: 10.1080/17421772.2016.1190462.

Herwartz, H., Tafenau, E., & Schneider, F. (2015). One Share Fits All? Regional Variations in the Extent of the Shadow Economy in Europe. *Regional Studies*, *49*(9), 1575–1587. doi: 10.1080/00343404.2013.848034.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. doi: 10.1080/01621459.1952.10483446.

Jonasson, E. (2012). Government Effectiveness and Regional Variation in Informal Employment. *Journal of Development Studies*, *48*(4), 481–497. doi: 10.1080/00220388.2011.615922.

Kelly, P. (2011). Space, Place and Labour. In A. Leyshon, R. Lee, L. McDowell, & P. Sunley (Eds.), *Economic geography.* London: Sage.

Kucera, D., & Xenogiani, T. (2009). Persisting Informal Employment: what explains it? In J. Jütting & J. de Laiglesia (Eds.), *Is informal normal? towards more and better jobs in*

developing countries (pp. 63–88). Paris: OECD Publishing. doi: 10.1787/9789264059245-en.

Łapiński, K., Peterlik, M., & Wyżnikiewicz, B. (2015). *Szara strefa w polskiej gospodarce w 2015 roku [The shadow economy in the Polish economy in 2015]* (Tech. Rep.). Warszawa: Instytut Badań nad Gospodarką Rynkową. Retrieved from `http://www.ibngr.pl/content/download/1635/17099/file/Szara%20Strefa%20-%C2%A0Raport%20IBnGR.pdf`.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25. doi: 10.18637/jss.v063.i19.

Marelli, E., Patuelli, R., & Signorelli, M. (2012). Regional unemployment in the EU before and after the global crisis. *Post-Communist Economies*, *24*(2), 155–175. doi: 10.1080/14631377.2012.675153.

Newell, A., & Pastore, F. (2006). Regional Unemployment and Industrial Restructuring in Poland. *Eastern European Economics*, *44*(3), 5–28. doi: 10.2753/EEE0012-8755440301.

Nikulin, D., & Sobiechowska-Ziegert, A. (in press). Informal work in Poland -– a regional approach. *Papers in Regional Science*. doi: 10.1111/pirs.12306.

Novotný, J., & Nosek, V. (2012). Comparison of regional inequality in unemployment among four Central European countries: an inferential approach. *Letters in Spatial and Resource Sciences*, *5*(95), 95–101. doi: 10.1007/s12076-011-0071-y.

Patuelli, R., Schanne, N., Griffith, D. A., & Nijkamp, P. (2012). Persistence of regional unemployment: Application of a spatial filtering approach to local labor markets in Germany. *Journal of Regional Science*, *52*(2), 300–323. doi: 10.1111/j.1467-9787.2012.00759.x.

Perry, G. E., Maloney, W. F., Arias, O. S., Fajnzylber, P., & Saavedra-Chanduvi, Jaime Mason, A. D. (2007). *Exit and Exclusion*. Washington D.C.: The Word Bank.

Polese, A., & Morris, J. (2015). My Name Is Legion. The Resilience and Endurance of Informality beyond, or in spite of, the State. In A. Polese & J. Morris (Eds.), *Informal economies in post-socialist spaces. practices, institutions and networks* (pp. 1–12). Palgrave Macmillan. doi: 10.1057/9781137483072.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145–1165. doi: 10.1177/0962280216660421.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. doi: 10.1093/biomet/63.3.581.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, *71*, 319–392. doi: 10.1111/j.1467-9868.2008.00700.x.

Schneider, F., & Enste, D. H. (2000). Shadow Economies : Size , Causes , and Consequences. *Journal of Economic Literature*, *XXXVIII*(March), 77–114. doi: 10.1257/jel.38.1.77.

Tafenau, E., Herwartz, H., & Schneider, F. (2010). Regional Estimates of the Shadow Economy in Europe. *International Economic Journal*, *24*(4), 629–636. doi: 10.1080/10168737.2010.526010.

Torgler, B., & Schneider, F. (2007). *Shadow Economy, Tax Morale, Governance and Institutional Quality: A Panel Analysis* (Discussion Paper No. 2563). Bonn: IZA. Retrieved from `http://ftp.iza.org/dp2563.pdf`

Tyrowicz, J., & Cichocki, S. (2011). Employed unemployed? On shadow employment in transition. *Empirica*, *38*(2), 259–281. doi: 10.1007/s10663-010-9135-2.

Venn, D. (2008). Measuring Informal Employment in OECD Countries. In *Measuring informal employment in developed countries* (pp. 1–4). Cambridge, MA: John F. Kennedy School of Government. Retrieved from `http://www.wiego.org/sites/default/files/publications/files/Venn{_}OECD{_}Countries.pdf`

Williams, C. C. (2015a). Evaluating cross-national variations in envelope wage payments in East-Central Europe. *Economic and Industrial Democracy*, *36*(2), 283–303. doi: 10.1177/0143831X13505120.

Williams, C. C. (2015b). Explaining Cross-National Variations in the Informalisation of Employment. Some lessons from Central and Eastern Europe. *European Societies*, *17*(4), 492–512. doi: 10.1080/14616696.2015.1051073.

Williams, C. C., & Horodnic, I. (2015a). Are marginalised populations more likely to engage in undeclared work in the nordic countries? *Sociological Research Online*, *20*(3). doi: 10.5153/sro.3719.

Williams, C. C., & Horodnic, I. (2015b). Marginalisation and participation in the informal economy in Central and Eastern European nations. *Post-Communist Economies*, *27*(2). doi: 10.1080/14631377.2015.1026686.

Williams, C. C., & Horodnic, I. A. (2015c). Self-employment, the informal economy and the marginalisation thesis: Some evidence from the European Union. *International Journal of Entrepreneurial Behavior & Research*, *21*(2), 224–242. doi: http://dx.doi.org/10.1108/09564230910978511.

Table 1. Direct estimates of share of different categories of participation in labour market in the last 12 months in Poland between 2010 and 2014 based on the BKL survey

Category		2010	2011	2012	2013	2014
Employed	No	54.6	54.8	53.8	53.6	53.1
	Yes	45.4	45.2	46.2	46.4	46.9
	Open-ended contract	–	–	–	20.0	24.1
	Fixed term contract	–	–	–	80.0	75.9
Informal work	No	95.3	95.9	96.2	96.2	96.0
	Yes	4.7	4.1	3.8	3.8	4.0
Self-employed	No	88.8	88.1	89.2	88.7	88.5
	Yes	11.2	11.9	10.8	11.3	11.5
Agricultural activity	No	94.2	94.7	95.7	95.5	95.2
for its own use	Yes	5.7	5.3	4.3	4.5	4.8
	N/A	0.1	–	–	–	
Contract for a specific task	No	93.3	93.2	92.9	92.8	92.8
or commission agreement	Yes	6.7	6.8	7.1	7.2	7.2

Table 2. Descriptive statistics of direct estimates of the share of informal workers at NUTS 3 level between 2010 and 2014 based on the BKL survey

Year	Min	Q1	Median	Mean	Q3	Max
2010	0.0	3.3	4.3	4.8	6.0	17.0
2011	0.0	2.5	4.0	4.2	5.4	11.2
2012	0.0	2.3	3.6	3.9	5.3	10.3
2013	0.9	2.8	3.6	3.8	4.9	8.2
2014	0.6	2.9	3.9	4.1	4.9	10.7

Table 3. Comparison of the estimated models

Model	Fixed effects	Mixed Effects	BESAG	BYM
Marginal LL	-13 522.52	-13 431.96	-13 466.71	-13 403.76
DIC	26 788.76	26 529.45	26 526.64	26 526.36
WAIC	26 789.18	26 529.77	26 527.19	26 526.90
Hyperparameters – Median and CI 95% in brackets				
σ_u (i.i.d)	–	0.30 (0.24, 0.38)	–	0.02 (0.01, 0.15)
σ_v (spatial)	–	–	0.43 (0.33, 0.58)	0.41 (0.30, 0.57)

Table 4. Estimated parameters (linear and odds ratio) for logistic regression with BESAG random effect

	Beta				Odds ratio			
Parameter	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
Intercept	-2.83	0.07	-2.97	-2.69	0.06	0.00	0.05	0.07
Trend	-0.07	0.02	-0.12	-0.03	0.93	0.02	0.89	0.97
Register unempl >= 12m	0.01	0.01	-0.01	0.02	1.01	0.01	0.99	1.02
Female	-0.60	0.07	-0.73	-0.47	0.55	0.04	0.48	0.62
Children = Yes	-0.35	0.04	-0.43	-0.26	0.71	0.03	0.65	0.77
Part-time worker = Yes	1.89	0.06	1.76	2.01	6.60	0.41	5.84	7.43
ILO (ref = 'Working')								
Unemployed = Yes	1.62	0.05	1.52	1.71	5.04	0.24	4.58	5.54
Inactive = Yes	0.55	0.05	0.45	0.65	1.74	0.09	1.57	1.92
Age (ref = '18-24')								
25-34	0.09	0.06	-0.04	0.21	1.09	0.07	0.96	1.23
35-44	-0.12	0.07	-0.27	0.02	0.89	0.07	0.77	1.02
45-54	-0.24	0.07	-0.38	-0.10	0.79	0.06	0.69	0.91
55-59/64	-0.90	0.08	-1.06	-0.73	0.41	0.03	0.35	0.48
Education (ref = 'Primary')								
Basic vocational	-0.31	0.05	-0.41	-0.21	0.74	0.04	0.67	0.81
Secondary	-0.49	0.05	-0.58	-0.40	0.61	0.03	0.56	0.67
Tertiary	-0.85	0.07	-0.99	-0.71	0.43	0.03	0.37	0.49
Locality (ref = 'Rural')								
Urban (<= 50k)	0.08	0.05	-0.01	0.17	1.09	0.05	0.99	1.19
Urban (50-100k)	0.25	0.07	0.12	0.39	1.29	0.09	1.12	1.47
Urban (100-200k)	0.34	0.08	0.19	0.48	1.40	0.11	1.21	1.62
Urban (200-500k)	0.23	0.08	0.06	0.39	1.26	0.10	1.06	1.48
Urban (> 500k)	0.29	0.20	-0.11	0.69	1.37	0.28	0.90	1.98
Sex-Age interaction								
Females 25-34	-0.36	0.10	-0.56	-0.16	0.70	0.07	0.57	0.85
Females 35-44	-0.23	0.11	-0.44	-0.01	0.80	0.09	0.64	0.99
Females 45-54	-0.40	0.11	-0.62	-0.18	0.67	0.07	0.54	0.83
Females 55-59/64	-0.16	0.16	-0.48	0.14	0.86	0.13	0.62	1.15

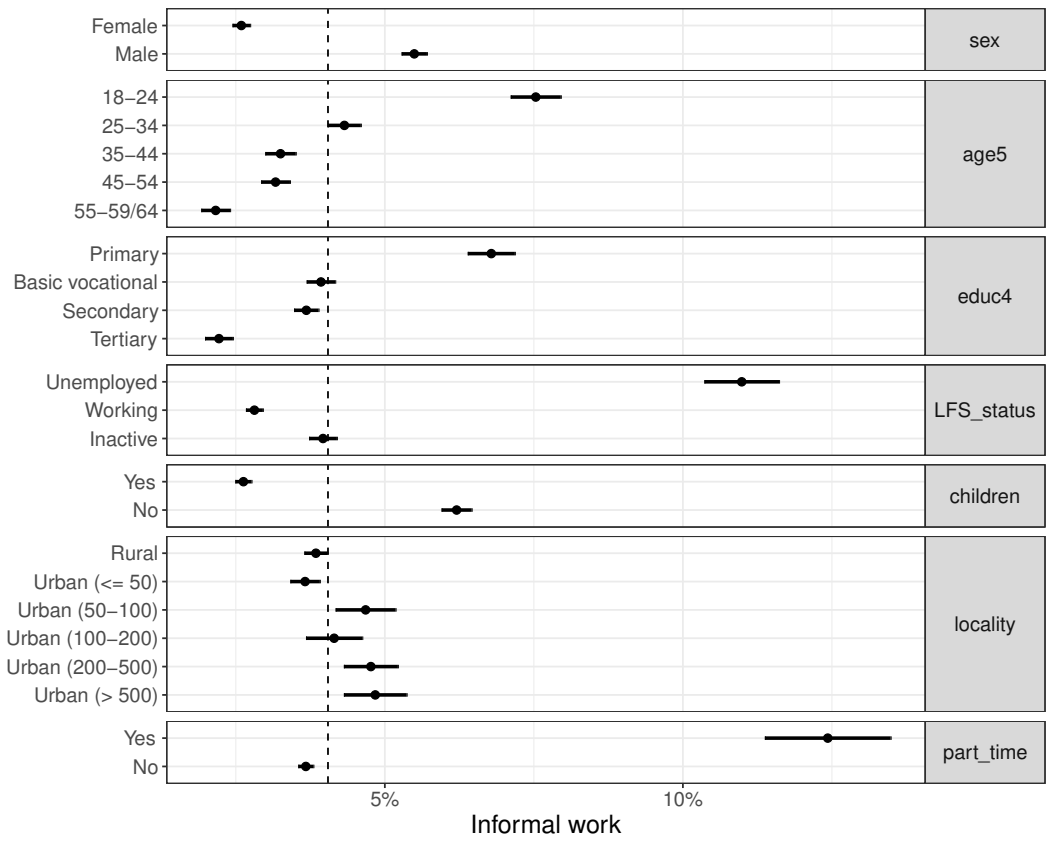


Figure 1. Exploratory analysis of the share of informal workers classified by demographic variables. Black dots denote the proportion of respondents in each group who worked informally. Black lines denote 95% confidence intervals for proportions. The dashed line denotes the overall proportion of informal workers across all subgroups between 2010 and 2014. Weights were not included in the calculations.

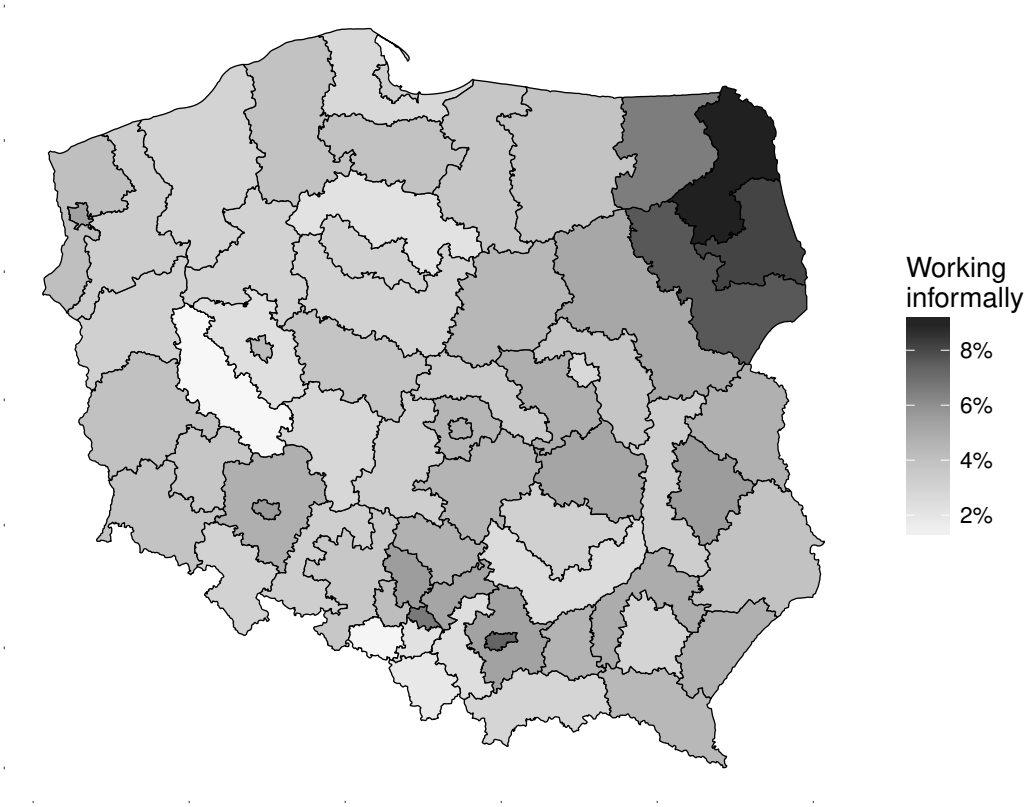


Figure 2. Spatial analysis of the share of informal workers between 2010 and 2014 (NUTS 3) based on polled samples

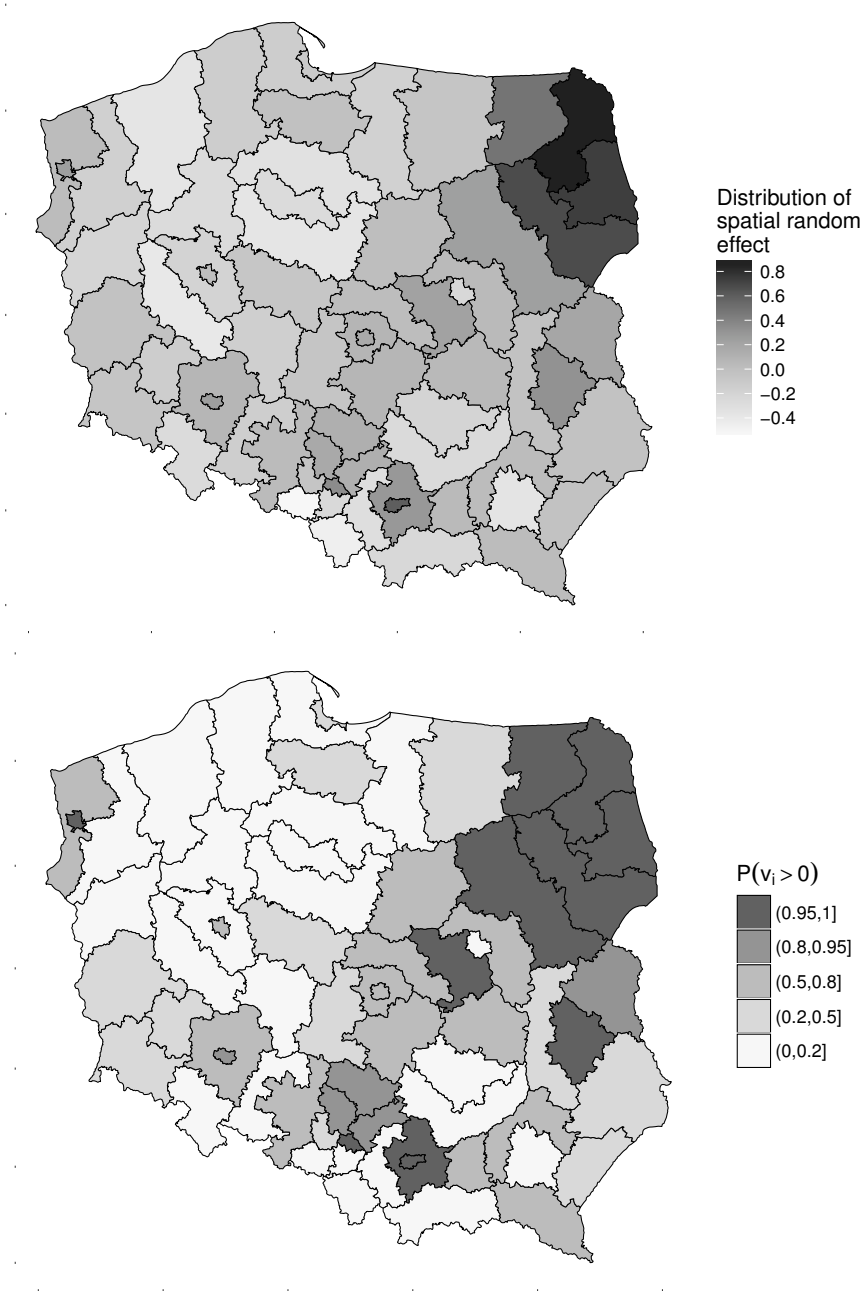


Figure 3. Distribution of the posterior median for the spatial random effect (upper figure, on the natural scale) and the probability that the posterior median of the spatial random effect is grater than 0 (lower figure). Estimates based on the BYM model. Map presents NUTS 3 units.

Online supplementary materials for the paper *Informal employment in Poland: an empirical spatial analysis* (Beręsewicz and Nikulin, 2018)

Appendix A. Detailed information about the survey

In this study we use unit-level data from the national annual BKL survey. It was conducted by the Polish Agency for Enterprise Development in cooperation with the Jagiellonian University. The first round was carried out in the fourth quarter of 2010, the subsequent ones, 2011, 2012, 2013, and 2014 were conducted during the second quarter of each year. The aim of the survey was to provide systematic data about the structure of competences available in the labour market in Poland. The survey consisted of four core research modules with additional surveys focusing on specific topics. The core modules included a survey of employers, job offers, persons of working age, and educational institutions. Additional surveys focused on pupils (2010 and 2013), students (2010 and 2013) and unemployed (2010). All field work was conducted by a private company Kantar Millward Brown, but the data are freely available online at `https://bkl.parp.gov.pl/dane.html`.

The questionnaire consisted of several parts including questions on employment history, competences or employment. In the part on employment history, the question regarding unregistered work was stated as follows *N0. Have you worked on the basis of an informal agreement, e.g. a verbal agreement, in the past 12 months?*. It was a separate question, in addition to questions regarding agricultural activities for own use (*R0. Have you conducted agricultural activities for own use in the past 12 months?* and performing unpaid work in a family business or farm (*C0. Have you done unpaid work in a family business or farm in the past 12 months?*). The question about an informal agreement was then followed by more detailed questions concerning the character of this work – whether it was in line with the respondent's educational background and the reasons for undertaking such work: (1) too high taxes and charges; (2) cumbersome formalities; (3) reluctance of other parties to sign a formal agreement; (4) the lack of formal qualifications; (5) performing work for family or friends; (6) being registered at the District Employment Agency; (7) just additional work, an odd job; (8) other reasons. The negative answer (Yes/No) to the question was used as an independent variable in our study. For replication purposes we provide all calculations in the supplementary materials and in Appendix A.

The study is based on a probability sample from a statistical population defined as persons of working age, that is aged 18-59 and 18-64 for women and men respectively. The sample was selected from the PESEL register (Universal Electronic System for Registration of the Population) based on a two-stage sampling scheme. In the first stage, municipalities were sampled with replacement stratified by subregion (NUTS 3 level, 66 units in total) and 9 location classes with probability proportional to the number of residents (in rural municipalities proportional to the population aged 18-59/65). In the second stage, simple random samples of persons stratified by sex and 5 age groups (18–29, 30-39, 40-49, 50-59/64) were selected. Within each municipality a cluster of 10 persons was selected and because PSUs were sampled with replacement, the initial sample size in each PSU was a multiple of 10. However, it should be noted that the PESEL register is not error-free and there is some under-coverage, particularly

in the case of persons aged 18–29, who are characterised by high mobility (Józefowski & Rynarzewska-Pietrzak, 2010).

The final weights were calculated based on a combination of design weights, return rate within stratification groups and post-stratification to known population totals defined by sex, 5 age groups and provinces (160 strata in total). Appendix A1 contains detailed information on sample size and non-response between 2010 and 2014.

In the analysis we focused on the NUTS 3 level (subregions) and not on the LAU-1 level for the following reasons. For one thing, the NUTS 3 level was included in the sampling scheme, hence all levels are present throughout the reference period. Some subregions are also the biggest cities in Poland (e.g. Warsaw, Poznań, Kraków, Wrocław). The sample size was not sufficient to provide reliable direct estimates at LAU-1 level (380 districts in total). Finally, not all districts were observed in the whole reference period and the sample size at LAU-1 varied from 1 to 377 with a median of 36 persons.

Not all variables were free from errors. The target variable, called `informal`, contained 7 missing values (for 2010 and 2011), which were removed prior to the modelling procedure. Variables `age5`, `children` and `marital` contained 108, 29 and 440 missing values respectively. In the case of `marital`, the majority of missing values were due to refusals. To deal with this problem we imputed missing values using the k-nearest neighbour algorithm based on the Gower Distance, which is implemented in the VIM package (Kowarik & Templ, 2016), assuming that data are missing at random (do not depend on the imputed variable itself).

Table A1. Basic statistics about the BKL survey administered between 2010 and 2014

Year	2010	2011	2012	2013	2014
CAPI	–	84%	77%	92.5%	90.7%
PAPI	–	16%	23%	7.5%	9.3%
Expected sample size	32 000	32 000	32 000	32 000	32 000
Realized sample size	17 899	17 780	17 600	17 600	17 674
Return rate	55.9%	55.6%	55.0%	55.0%	55.2%

Table A2. Exploratory analysis of the share of informal workers classified by demographic variables [in %]

Variable	level	Share	CI		Sampe size
Age	18-24	7.53	7.11	7.95	1142
	25-34	4.32	4.05	4.60	901
	35-44	3.25	2.99	3.51	600
	45-54	3.17	2.93	3.41	640
	55-59/64	2.16	1.92	2.40	300
Children	No	6.20	5.95	6.46	2179
	Yes	2.63	2.49	2.76	1404
Education	Primary	6.79	6.39	7.18	1056
	Secondary	3.68	3.48	3.89	1196
	Tertiary	2.22	1.98	2.45	338
	Basic vocational	3.93	3.69	4.17	993
LFS status	Unemployed	10.99	10.36	11.61	1052
	Inactive	3.96	3.73	4.20	1068
	Working	2.81	2.67	2.95	1463
Locality	Rural	3.84	3.65	4.04	1459
	Urban (<= 50k)	3.66	3.41	3.91	809
	Urban (50k-100k)	4.68	4.17	5.18	317
	Urban (100k-200k)	4.15	3.68	4.62	284
	Urban (200k-500k)	4.77	4.31	5.22	405
	Urban (> 500k)	4.84	4.31	5.37	309
Sex	Female	2.59	2.44	2.74	1146
	Male	5.49	5.28	5.71	2437
Part time	Yes	12.43	11.38	13.49	467
	No	3.67	3.55	3.80	3116

Appendix B. Using weights for estimation

The discussion about the use of sampling weights is not new and remains the subject of ongoing research (Gelman, 2007; Pfeffermann, 2011). However, owing to the lack of weights from the first stage of sampling we were not able to verify correctly how informative the sampling scheme was in the context of mixed models (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998; Pfeffermann & Sverchkov, 2007). Therefore, we decided to estimate two logistic regression models (fixed and mixed effects with an unstructured random effect) with and without weights. To identify the impact of the registered unemployment over 12 months rate, we also calculated these two models with and without accounting for this variable. We used `survey` and `lme4` packages. Estimation results are presented in Table B1 and B2. To test the ignorability of the sampling scheme for GLM we applied a Hausman-like test statistic discussed by Pfeffermann (1993, p. 324–325), which is based on a comparison of estimated parameters taking into account their standard errors. However, this test cannot be applied directly for GLMM, because these models consider complex variance structures. That is why we decided to use some ideas from small area estimation, in particular unit-level models. Burgard, Münnich, and Zimmermann (2014, eq. 9) show a method to verify informativeness of sampling design by extending the unweighted EBLUP under the unit-level mixed model is by augmenting the design matrix by the design weights. Coefficient associated with weight is used to detect impact of weights on estimation. Test results are presented in Table B3 for fixed and mixed models with and without the registered unemployment over 12m.

The results indicate that sampling weights significantly change the vector of GLM parameters only for the case when rate of registered unemployed over 12 m is used. There is still a difference with regard to weighted and unweighted mixed models (e.g. Age 55-59/65, Females 35-44, Secondary education). Nonetheless, coefficient associated with weight, as suggested by Burgard et al. (2014, eq. 9), is not significant which indicates that sampling design is not informative. That is why we finally decided not to include sampling weights in the proposed models.

Table B1. Estimates of parameters from fixed effects logistic regression with and without sampling weights (standard errors are given in brackets)

Parameter	GLM (unweighted)	GLM (weighted)
(Intercept)	-2.79 (0.07)***	-2.76 (0.08)***
Trend	-0.12 (0.02)***	-0.12 (0.02)***
Register unempl >=12m	0.02 (0.00)***	0.02 (0.00)***
Female	-0.59 (0.07)***	-0.60 (0.08)***
Children = Yes	-0.34 (0.04)***	-0.37 (0.05)***
Part-time worker = Yes	1.92 (0.06)***	1.94 (0.07)***
ILO (ref = 'Working')		
Unemployed = Yes	1.61 (0.05)***	1.61 (0.06)***
Inactive = Yes	0.55 (0.05)***	0.53 (0.07)***
Age (ref = '18-24')		
25-34	0.07 (0.06)	0.07 (0.08)
35-44	-0.13 (0.07)	-0.09 (0.09)
45-54	-0.23 (0.07)***	-0.24 (0.08)**
55-59/64	-0.90 (0.08)***	-0.97 (0.10)***
Education (ref = 'Primary')		
Basic vocational	-0.32 (0.05)***	-0.32 (0.06)***
Secondary	-0.48 (0.05)***	-0.54 (0.06)***
Tertiary	-0.82 (0.07)***	-0.83 (0.09)***
Locality (ref = 'Rural')		
Urban (<= 50k)	0.10 (0.05)*	0.10 (0.06)
Urban (50-100k)	0.33 (0.07)***	0.34 (0.08)***
Urban (100-200k)	0.27 (0.07)***	0.30 (0.10)**
Urban (200-500k)	0.41 (0.06)***	0.42 (0.09)***
Urban (> 500k)	0.45 (0.07)***	0.42 (0.09)***
Sex-Age interaction		
Females 25-34	-0.35 (0.10)***	-0.37 (0.12)**
Females 35-44	-0.23 (0.11)*	-0.31 (0.12)*
Females 45-54	-0.42 (0.11)***	-0.30 (0.13)*
Females 55-59/64	-0.17 (0.16)	-0.15 (0.17)
AIC	26788.75	
BIC	27014.14	
Log Likelihood	-13370.37	
Deviance	26740.75	26769.17
Num. obs.	88553	88553
Dispersion		0.99

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B2. Estimates of parameters from the mixed effects logistic regression with and without sampling weights (standard errors are given in brackets)

Parameter	GLMM (unweighted)	GLMM (weighted)
(Intercept)	-2.82 (0.08)***	-2.76 (0.08)***
Trend	-0.10 (0.02)***	-0.12 (0.02)***
Register unempl >=12m	0.02 (0.01)*	0.02 (0.01)**
Female	-0.60 (0.07)***	-0.60 (0.07)***
Children = Yes	-0.35 (0.04)***	-0.37 (0.04)***
Part-time worker = Yes	1.89 (0.06)***	1.92 (0.06)***
ILO (ref = 'Working')		
Unemployed = Yes	1.62 (0.05)***	1.60 (0.05)***
Inactive = Yes	0.56 (0.05)***	0.53 (0.05)***
Age (ref = '18-24')		
25-34	0.09 (0.06)	0.09 (0.06)
35-44	-0.12 (0.07)	-0.08 (0.07)
45-54	-0.23 (0.07)**	-0.24 (0.07)***
55-59/64	-0.89 (0.08)***	-0.96 (0.09)***
Education (ref = 'Primary')		
Basic vocational	-0.31 (0.05)***	-0.31 (0.05)***
Secondary	-0.48 (0.05)***	-0.54 (0.05)***
Tertiary	-0.84 (0.07)***	-0.85 (0.07)***
Locality (ref = 'Rural')		
Urban (<= 50k)	0.09 (0.05)	0.11 (0.05)*
Urban (50-100k)	0.25 (0.07)***	0.27 (0.07)***
Urban (100-200k)	0.33 (0.08)***	0.32 (0.08)***
Urban (200-500k)	0.22 (0.08)**	0.25 (0.08)**
Urban (> 500k)	0.47 (0.16)**	0.47 (0.15)**
Sex-Age interaction		
Females 25-34	-0.36 (0.10)***	-0.37 (0.10)***
Females 35-44	-0.23 (0.11)*	-0.31 (0.11)**
Females 45-54	-0.41 (0.11)***	-0.29 (0.11)**
Females 55-59/64	-0.17 (0.16)	-0.15 (0.16)
AIC	26589.75	24652.40
BIC	26824.53	24887.19
Log Likelihood	-13269.87	-12301.20
Num. obs.	88553	88553
Num. groups: podregion	66	66
Var: podregion (Intercept)	0.09	0.08

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table B3. Table contains results of the Hausman-like test proposed by Pfeffermann (1993, eq. 4.3) and coefficient test for weight proposed by Burgard et al. (2014, eq. 9) for models built with and without sampling weights

Model	Test statistic	p-value
Models with Register Unemployment >= 12m		
Fixed effects	82.6	<0.001
Mixed effect (weight as covariate)	0.067	0.24
Models without Register Unemployment >= 12m		
Fixed effects	14.9	0.89
Mixed effect (weight as covariate)	0.068	0.24

Appendix C. Comparison of fixed effects parameters point and 95% credible intervals estimates

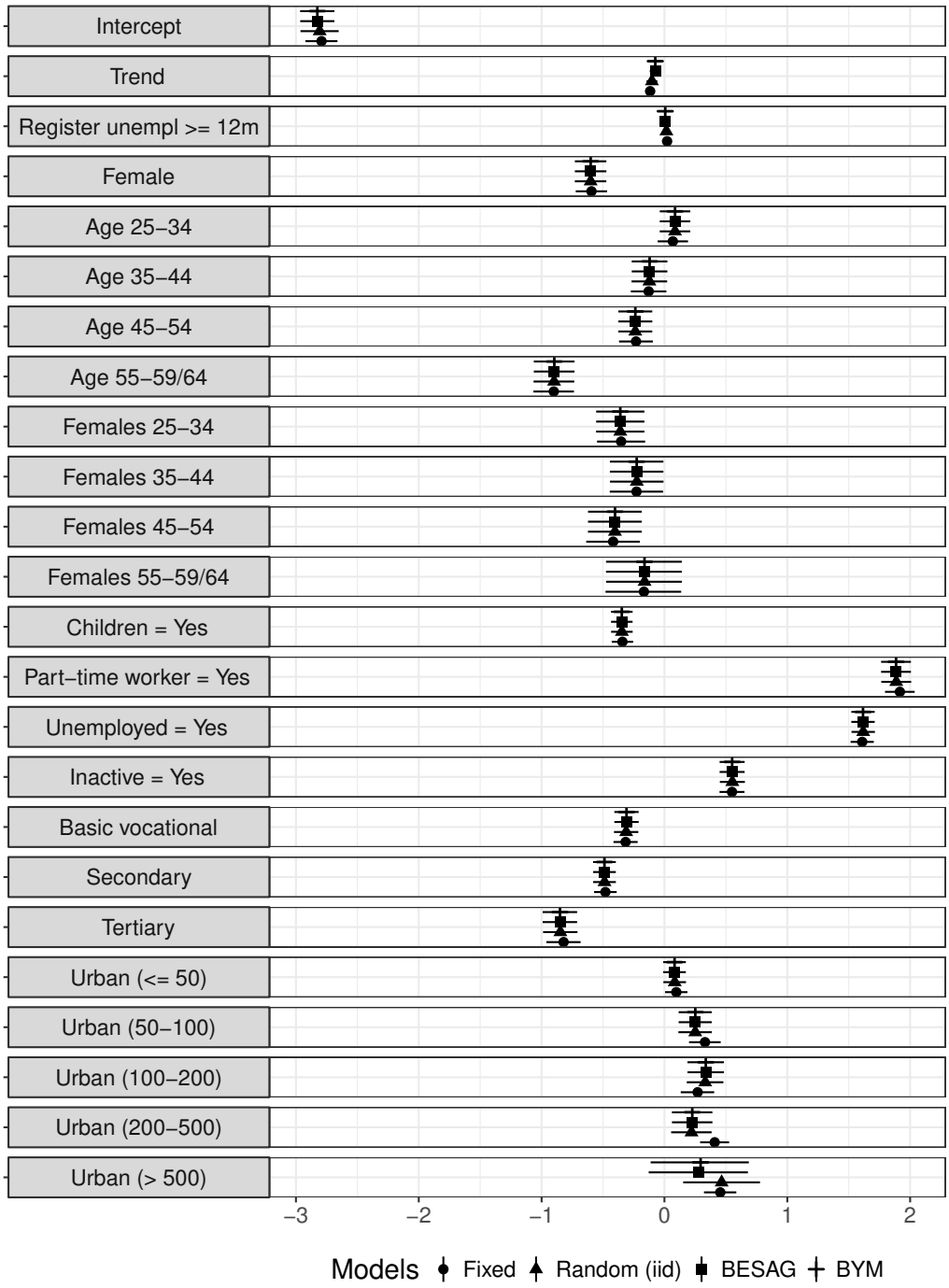


Figure C1. Comparison of fixed effects parameter point and 95% credible intervals estimates obtained from four models estimated in INLA.

Appendix D. Comparison of random effects point and 95% credible intervals estimates

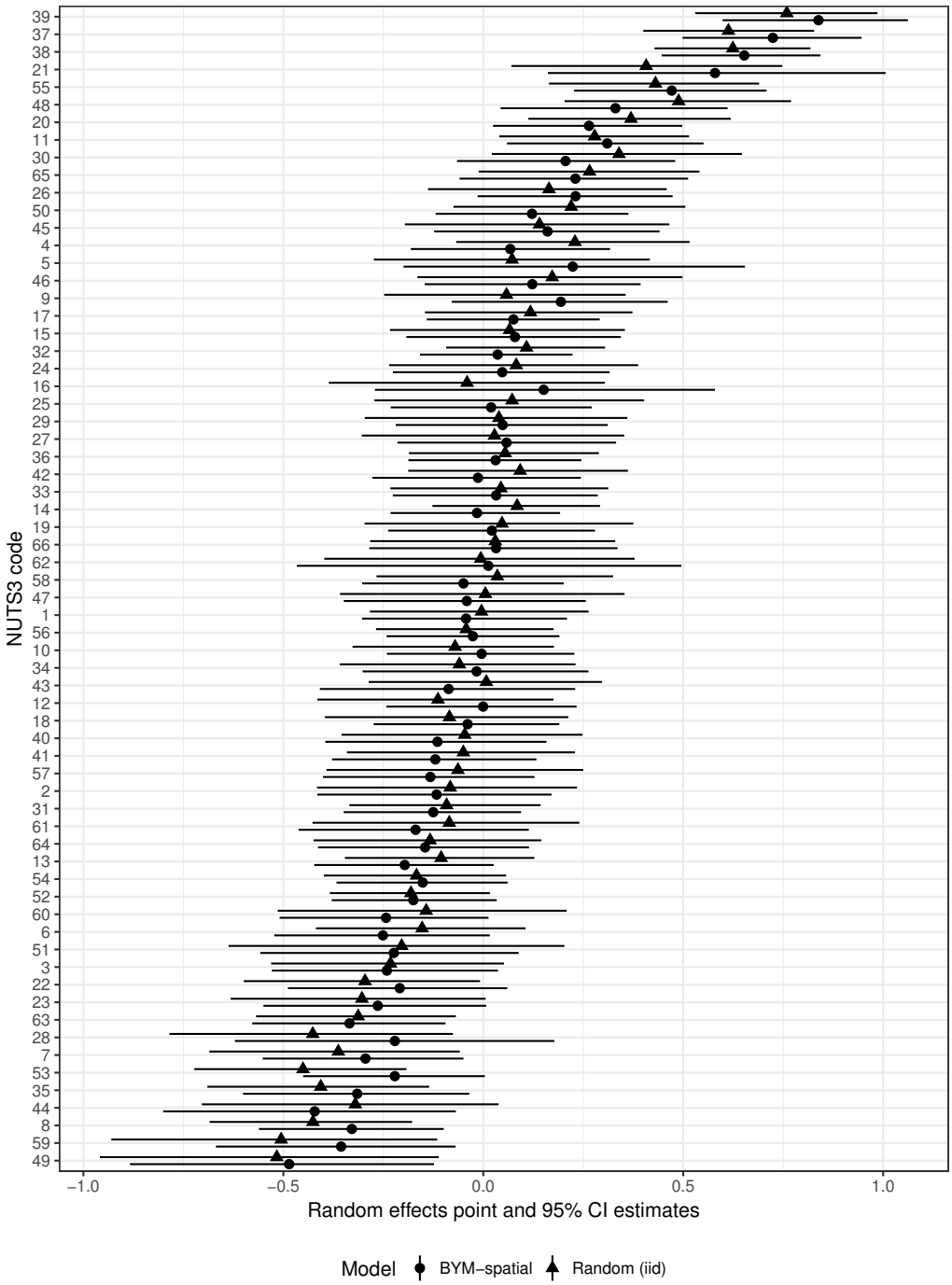


Figure D1. Comparison of random effects point and 95% credible intervals estimates from two models – random effect with iid effect and spatial random effect from BYM model

Appendix E. Distribution of the propensity for informal work

Figure E1 presents the distribution of the propensity for informal work among people of working age. The distribution is highly skewed with a minimum of 0.0017, a maximum of 0.5128, a mean of 0.0404 and a median of 0.0241. The shape of the propensity distribution indicates that outliers are present in the sample. This calls for the use of robust methods or methods based on quantiles in further analysis. A person with the highest propensity for informal is a young (18-24) male with primary education, who works part-time and has no children. In contrast, a person with the lowest propensity is an older (55-59) female with tertiary education who lives in a rural area and has children.

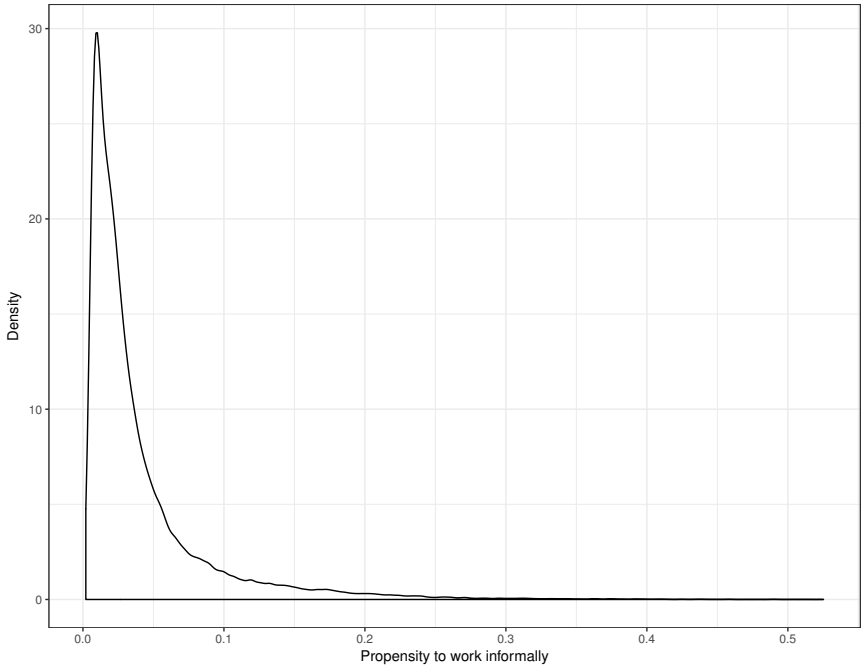
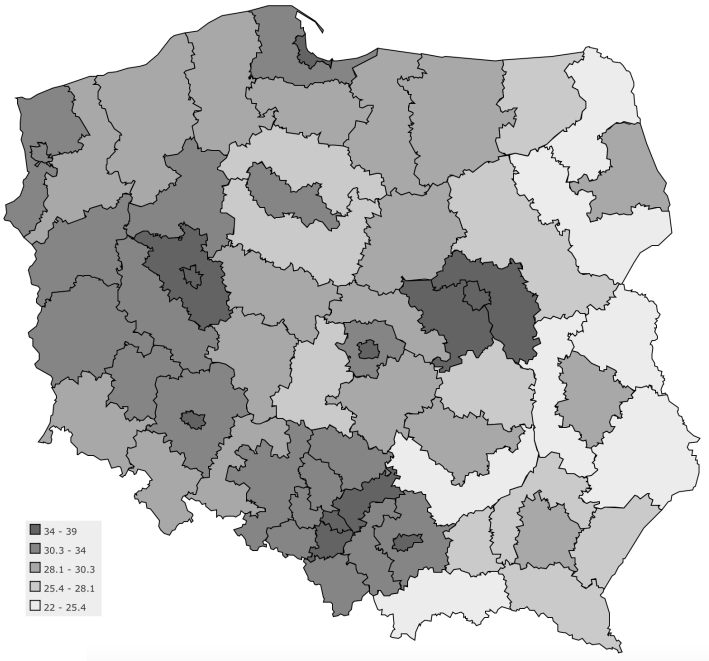


Figure E1. The distribution of the median of posterior estimates of the propensity for informal work in the working age population in Poland between 2010 and 2014.

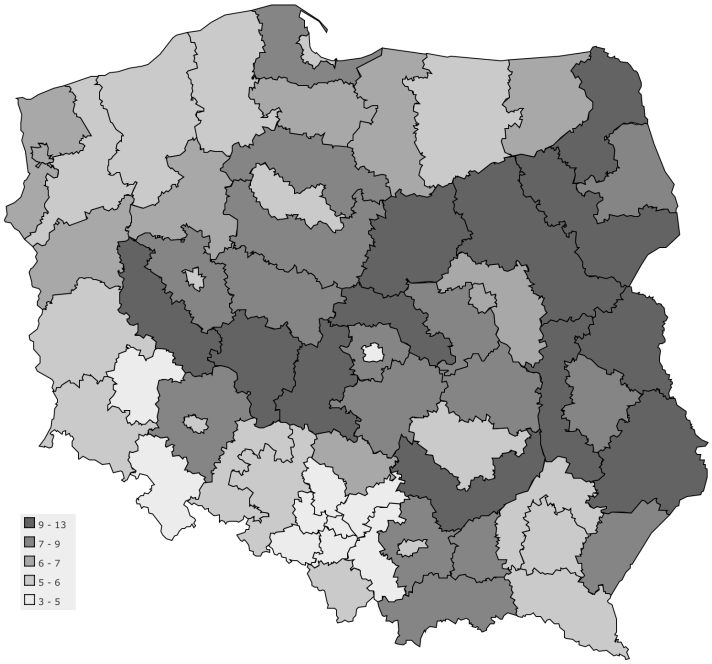
Appendix F. The Nomenclature of Territorial Units for Statistical Purposes in Poland

Figure F1. The Nomenclature of Territorial Units for Statistical Purposes in Poland. In Poland we have 6 NUTS 1, 16 NUTS 2, 72 NUTS 3 level units. NUTS 3 level – subregions (groups of powiats). Names refer to NUTS 3 units, solid line to NUTS 2 level and grey lines within NUTS 3 level refer to LAU 1 units. Until 31.12.2014 there were 66 units, from 01.01.2015 it was increased to 72 units. Source: Central Statistical Office.

Appendix G. Regional situation at NUTS 3 level

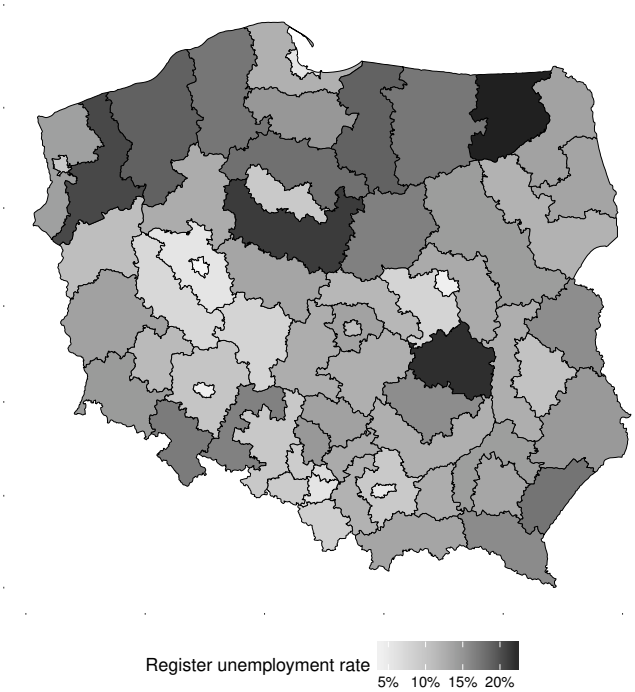


(a) Share of paid employees in total employment

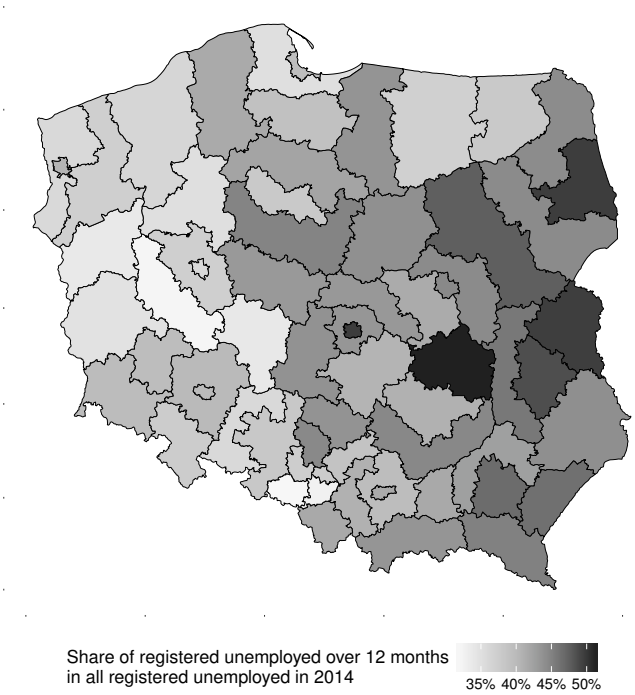


(b) Share of self-employed in total employment

Figure G1. The National Census of Population and Housing 2011 – employment status at NUTS 3 level. Source: Central Statistical Office `http://stat.gov.pl/en/national-census/national-census-of-population-and-housing-2011/`



(a) Registered unemployment rate



(b) Share of registered unemployed persons out of job for longer than 1 year in all registered unemployed

Figure G2. Registered unemployment statistics in 2014 at NUTS 3 level. Source: Central Statistical Office.

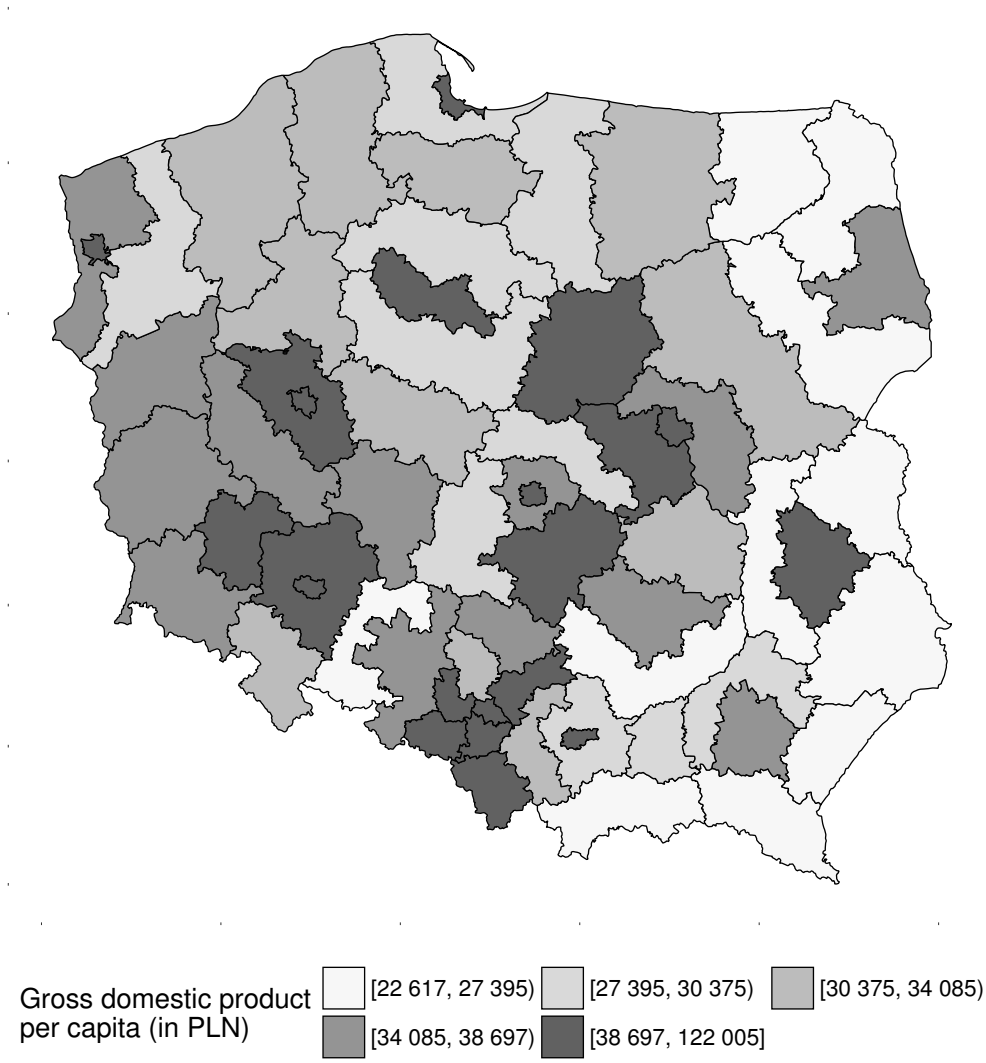


Figure G3. Gross domestic product per capita (in PLN, current prices). Category classes based on quantiles in 2014.

References

Burgard, J. P., Münnich, R., & Zimmermann, T. (2014). The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics*, *30*(4), 749–771. doi: 10.2478/jos-2014-0046.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, *22*(2), 153–164. doi: 10.1214/088342306000000691.

Józefowski, T., & Rynarzewska-Pietrzak, B. (2010). Ocena możliwości wykorzystania rejestru PESEL w spisie ludności [Assessment of Using PESEL Register in Population Census]. *Zeszyty Naukowe / Uniwersytet Ekonomiczny w Poznaniu*(149), 67–84. Retrieved from `http://yadda.icm.edu.pl/bazekon/element/bwmeta1.element.ekon-element-000171205903`

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, *74*(7), 1–16. doi: 10.18637/jss.v074.i07.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*, 317–337. doi: 10.2307/1403631.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*(12), 1–23. Retrieved from `http://www.statcan.gc.ca/pub/12-001-x/2011002/article/11602-eng.pdf`

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B*, *60*(1), 23–40. doi: 10.1111/1467-9868.00106.

Pfeffermann, D., & Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, *102*(480), 1427–1439. doi: 10.1198/016214507000001094.