Cena 15,43 zł
(VAT 8%)

Indeks 371262
ISSN 0033-2372

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 66

3

2019

GUS

Information for Authors

Przegląd Statystyczny (Statistical Review) publishes original research on theoretical and empirical topics in statistics, econometrics, mathematical economics, operational research, decision sciences and data analysis. The manuscripts considered for publication are ones which significantly contribute to the theoretical aspects of the aforementioned fields or shed new light on the practical applications of these aspects. Manuscripts reporting on important results of research projects are particularly welcome. Review papers, shorter papers reporting on major conferences in the field and reviews of seminal monographs are eligible for submission but only upon the Editor's request.

From May 1st, 2019 onwards the journal publishes articles in the English language. Any spelling style is acceptable as long as it is consistent within the manuscript.

All work should be submitted to the journal through the ICI Publishers Panel (https://editors.publisherspanel.com/pl.ici.ppanel-app-war/ppanel/index).

For the details of submission and editorial requirements please inspect https://ps.stat.gov.pl/ForAuthors.

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 66

3

2019

WARSZAWA 2019

ADRES REDAKCJI / EDITORIAL OFFICE
Uniwersytet Gdański, ul. Armii Krajowej 101, 81–824 Sopot

Redakcja językowa / Language editing: Wydział Czasopism Naukowych, Główny Urząd Statystyczny

Strona internetowa / Website: ps.stat.gov.pl

ISSN 0033-2372

e-ISSN 2657-9545

Indeks 371262

ZAKŁAD WYDAWNICTW STATYSTYCZNYCH
al. Niepodległości 208, 00-925 Warszawa, tel. 22 608 31 45.
Andrzej Paluchowski (redaktor techniczny), Katarzyna Szymańska (skład i łamanie)

Informacje w sprawie sprzedaży czasopisma tel.: 22 608 32 10, 22 608 38 10

Zam. 93/2019 — nakład 220 egz.

CONTENTS

REPORTS

DOI: 10.5604/01.3001.0013.9127

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Victor BYSTROV[1]

The observational equivalence of natural and unnatural rates of interest

Abstract. *The results of the study presented in this paper demonstrate that a structural model of the natural interest rate, which is consistent with the standard assumptions of the natural rate theory, admits an interpretable, observationally equivalent representation in which a redefined, 'unnatural' equilibrium rate is different from the natural rate in the original model. The alternative representation was obtained by an invertible transformation implemented in the minimal state-space form of the natural-rate model. The identification theory for state-space models is used in the paper to prove the observational equivalence of these two representations. In the alternative representation, the equilibrium interest rate fails to meet the assumption of the natural rate theory, because it depends on past demand shocks. The alternative model, being observationally equivalent, has different implications for the conduct of monetary policy. The problem of observational equivalence arises in relation to natural-rate models because of the inherent unobservability of the natural interest rate; a potential solution to this problem could be the augmentation* of *the information set which is used to identify and estimate the natural rate.*

Keywords: natural rate of interest, state-space model, observational equivalence

JEL Classification: C32, C51, E43

1. INTRODUCTION

The aim of this paper is to demonstrate, using a structural model of the natural (equilibrium) interest rate, that it is possible to find its interpretable, observationally equivalent alternative. The alternative model allows a different interpretation of the equilibrium interest rate and has different implications for the conduct of the monetary policy.

In the 1976 paper, Thomas J. Sargent demonstrated that reduced-form models would not permit to determine the difference between the natural rate theory and its alternatives: 'there are always alternative ways of writing the reduced form, one being observationally equivalent with the other, so that each is equally valid in the estimation period.' (Sargent, 1976, p. 631). Therefore, the rational expectations econometrics, using cross-equation parameter restrictions, has been developed in order to solve the problem of observational equivalence. However, parameter restrictions leave unresolved what Alan J. Preston (1978) called the model identification problem, referring to the fact that there are many models with identified parameters that provide the same fit to the data.

[1] University of Łódź, Faculty of Economics and Sociology, 3/5 POW Street, 90–255 Łódź, Poland, e-mail: victor.bystrov@uni.lodz.pl, ORCID: https://orcid.org/0000-0003-0980-2790.

The natural (equilibrium) rate of interest can be defined as a real rate of interest consistent with real output equalling its potential level in the absence of transitory shocks to demand (Williams, 2003). The potential output is the level of output consistent with the dynamic general equilibrium in the absence of nominal rigidities. Structural models of the natural rate are based on two assumptions, namely, the natural rate of interest is independent of the output gap (difference between actual and potential levels of real output), and a positive output gap cannot be sustained without accelerating inflation.

The model-dependent natural rate of interest is commonly used in the analysis of the monetary policy (see Laubach and Williams, 2003 and 2016; Holston et al., 2017; Fries et al., 2018). The monetary policy stance is defined by the real rate gap (the difference between the measured real rate of interest and the natural rate): a positive real rate gap means a contractionary policy stance, while a negative real rate gap means an expansionary stance. A contractionary or expansionary policy stance is achieved by pursuing policies which vary the real rate of interest with respect to the natural rate. In natural-rate models, monetary policy is neutral with respect to the potential output and the natural rate of interest, thus its scope is limited to the variations of the output gap.

The identification and estimation of the natural rate is usually carried out within a state-space representation of the corresponding structural model where the natural rate is modeled as an unobservable state variable. Two state-space structures are defined to be observationally equivalent if they imply the same probability distribution (likelihood function) for observable variables, and a structure is said to be identifiable if there is no other observationally equivalent structure (Rothenberg, 1971).

A method of identification of state-space structures, developed by Wall (1987), employs a blend of control theory and econometrics. Given that the likelihood function of a linear dynamic system is completely determined by the first two moments, two state-space representations are observationally equivalent if they give rise to the same first two moments of observable variables. Using this property of linear dynamic systems and the concept of minimal representation (a type of representation that includes no redundant state variables), Wall (1987) defined a class of observationally equivalent state-space structures and gave an operational criterion of observational equivalence.

The identification and estimation of state variables require the specification of initial states. For a state-space representation of a stationary multivariate process, initial states can be specified as functions of parameters describing that state-space representation. For a non-stationary process, on the other hand, the state-space representation should be augmented by a model for initial states (see Hamilton, 1994, or Durbin and Koopman, 2012).

A minimal representation, which includes no redundant state variables, guarantees that for given parameter matrices, a change of initial states will imply

a change of the likelihood value. For a non-minimal representation, the same value of the likelihood function can be obtained for different initial states and the same parameter matrices. Hence, for a non-minimal representation of a non-stationary process, different realizations of state variables (e.g. the natural rate of interest) can be obtained given the same parameter matrices and the same likelihood value. In other words, a non-minimal state-space representation of a non-stationary process admits an observationally equivalent representation which has the same parameter matrices and different values of state variables.

The natural-rate model considered in this paper is a modification of the widely-used model developed by Laubach and Williams (2003, 2016). The original Laubach-Williams model does not admit an irreducible state-space representation: the model specification used in that model requires the inclusion of redundant state variables (see Appendix). The model considered in this paper is consistent with the assumptions of the natural rate theory and admits an interpretable irreducible state-space representation. This representation is used to obtain an observationally equivalent model, where a redefined, 'unnatural' rate of interest depends on past output gaps, which is called hysteresis, and which involves the dependence of the equilibrium rate on the path the economy experiences towards the equilibrium.

The hysteresis hypothesis explains the fact that the estimates of natural interest rates in advanced industrial economies have been invariably low in the aftermath of the financial crisis of 2007–2008 (see Laubach and Williams, 2016; Holston et al., 2017; Fries et al., 2018). The low estimates of natural interest rates are often explained by persistent deviations from long-run trends ('headwinds', as coined by Yellen, 2015). In natural-rate models, these 'headwinds' come as components of natural interest rates that are exogenous with respect to the output gap. However, the association of recessions with low estimates of natural rates is consistent with the feedback from the output gap to natural rates transferred by "headwinds".

The model with hysteresis has different implications for the conduct of monetary policy: given the fact that the output gap depends on the stance of monetary policy, the equilibrium rate of interest, which depends on past output gaps, can also be affected by monetary policy. And because prolonged recession is likely to cause a downward shift in the equilibrium rate of interest, a more aggressive policy intervention is sometimes necessary in order to close the real rate gap.

The paper consists of four sections. Section 1 introduces the subject of the study, Section 2 provides a brief review of literature on the subject, Section 3 describes observationally equivalent irreducible state-space structures (in general terms), Section 4 demonstrates the observational equivalence of a natural-rate model and a model with hysteresis and Section 5 presents the conclusions of the study.

2. LITERATURE REVIEW

The concept of the natural interest rate, devised by Knut Wicksell (1898), has become popular in empirical research following the publication of Laubach and Williams (2003), in which a small semi-structural model was used to measure the natural rate of interest in the United States. Some modifications of this model were estimated for the Euro Area and other economies (Mésonnier and Renne, 2007; Garnier and Wilhelmsen, 2009; Holston et al., 2017). There are also modifications of the natural-rate model for the open-economy framework (Fries et al., 2018; Wynne and Zhang, 2018a), as well as the attempts to estimate the world natural rate of interest (Wynne and Zhang, 2018b; Kiley, 2019). Although various alternative approaches to the estimation of the natural interest rate have been proposed (Fiorentini et al., 2018; Grossman et al., 2019), the Laubach-Williams model and its modifications have become the most popular empirical tool for measuring the natural rate of interest, frequently cited by policy-makers (Yellen, 2015; Constancio, 2016).

Along with the increasing number of articles utilizing either the Laubach-Williams model or its modifications, there have also been a growing number of papers criticizing this approach. For example, the estimates of the natural rate of interest were found uncertain (Hamilton et al., 2016; Taylor and Wieland, 2016; Beyer and Wieland, 2017), as well as dependent on a priori assumptions concerning the structural relations between unobservable variables (Lewis and Vazquez-Grande, 2017).

Fiorentini et al. (2018) argue that the natural interest rate in the Laubach and Williams (2003) model is unobservable under certain conditions. The authors analyse a state-space representation of a simplified Laubach-Williams model, and demonstrate that state variables, including the determinants of the natural interest rate, are unobservable in two cases – either when the IS curve or the Phillips curve are flat. The unobservability implies that the natural rate is not uniquely identified by the model and the data. Fiorentini et al. (2018) propose a local-level model as an alternative to the Laubach-Williams model.

It can also be demonstrated that the original Laubach-Williams model is not consistent with the observability requirement either (see Appendix). Although a model that is inconsistent with the observability requirement can be transformed into a model that is consistent with that requirement, such transformation would result in a loss of the original economic interpretation. The modification of the Laubach-Williams model that is presented in this paper both fulfils the observability condition and retains the original economic interpretation. For such a model, there is a well-defined class of observationally equivalent models.

3. OBSERVATIONALLY-EQUIVALENT STATE-SPACE STRUCTURES

The state-space representation presented in this paper is

Transition Equation: $$\xi_t = F\xi_{t-1} + Gx_t + Qv_t, \quad (1)$$

Measurement Equation: $$y_t = H\xi_t, \quad (2)$$

where ξ_t is a $p \times 1$ vector of state variables, y_t is an $n \times 1$ vector of the observed explained variables, x_t is a $k \times 1$ vector of the observed exogenous variables, and v_t is a $q \times 1$ vector of structural shocks; F, G, Q and H are system matrices of dimensions $p \times p$, $p \times k$, $p \times q$ and $n \times p$, correspondingly.

The state-space representation (1)–(2) differs from the state-space representation used in Laubach and Williams (2003), where some dynamic relations between variables were included in the measurement equation. The representation (1)–(2) encompasses all the dynamic relations in the transition equation. Nevertheless, both the original Laubach-Williams model (see Appendix) and the modification considered in this paper admit the representation (1)–(2). This representation facilitates the analysis of observational equivalence without changing structural relations between variables. Because all shocks both in the original Laubach-Williams model and the modification considered in this paper determine the model dynamics, all these shocks appear in the transition equation (1). Although there are no measurement errors in the equation (2), the methodology, described below, would also apply if there were such errors.

The first two moments of explained variables y_t, $\mu_t = E\{y_t\}$ and $\Gamma(t,s) = E\{(y_t - E\{y_t\})(y_s - E\{y_s\})'\}$, are given by

$$\mu(t) = HF'\bar{\xi}_0 + H\sum_{s=1}^{t} F^{t-s}Gx_s,$$

$$\Gamma(t,s) = \begin{cases} HF^{t-s}P_sH' & \text{if } t > s \\ HP_tH' & \text{if } t = s, \\ HP_t(F^{s-t})'H' & \text{if } t < s \end{cases}$$

where $P_t = FP_{t-1}F' + QQ'$ is the covariance matrix of state variables; and $\bar{\xi}_0$ and P_0 are the initial conditions. If the eigenvalues of F are all inside the unit circle, then the process $\{\xi_t\}$ is stationary and the initial conditions are determined by the unconditional moments of this process: $\bar{\xi}_0 = E\{\xi_0\}$ and $P_0 = E\{(\xi_0 - E\{\xi_0\})(\xi_0 - E\{\xi_0\})'\}$. However, if some eigenvalues of F are on the unit circle, then the process $\{\xi_t\}$ is non-stationary and $\bar{\xi}_0$ can represent a guess as to the value of ξ_0 based on prior information, while P_0 measures the uncertainty associated with the guess (Hamilton, 1994).

Following Wall (1987), we say that the two state-space structures $S^{(i)} = \{F^{(i)}, G^{(i)}, Q^{(i)}, H^{(i)}\}$ $(i = 1,2)$ are observationally-equivalent if they produce the same first two moments of y_t. (This definition applies to minimal as well as non-minimal representations).

The state-space representation (1)–(2) is minimal if the dimension of the state-space cannot be reduced without a loss of information about responses of the explained variables y_t to the structural shocks v_t. If the representation (1)––(2) is not minimal, state variables and impulse responses are not uniquely identified and the model cannot be used for policy analysis.

A formal definition of a minimal structure uses the impulse response function $\Phi(h) = \sum_{j=0}^{h} HF^jQ$ and its weighting matrices $W(j) = HF^jQ$. A state-space structure is minimal if for any sequence $j = 0,1, \ldots, h$ such that $h \geq (p-1)$, the matrix

$$\begin{bmatrix} HQ \\ HFQ \\ \vdots \\ HF^hQ \end{bmatrix}$$

allows a full-rank decomposition $A(h)B(h)$, where $A(h)$ is an $(h \times n) \times p$ matrix of a full column rank and $B(h)$ is a $p \times q$ matrix of a full row rank (see, e.g., Youla, 1966).

The minimality test is based on checking the rank conditions for the weighting matrices of the structural shocks (Youla, 1966, Lemma 6):

Observability Condition: $$rank \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{p-1} \end{bmatrix} = p, \tag{3}$$

Controllability Condition: $$rank[Q\ FQ\ \ldots\ F^{p-1}Q] = p, \tag{4}$$

where p is the dimension of the state vector ξ_t.

The observability condition (3) implies that the state-space model (1)–(2) includes no state variables that cannot be inferred from observable variables. The controllability condition (4) implies that the state-space model includes no state variables independent of structural shocks. The observability and controllability are the necessary and sufficient conditions for the minimality of the system (1)–(2). These conditions guarantee a unique identification of state variables and enable the implementation of impulse-response analysis in the model.

The structures $S^{(1)} = \{F^{(1)}, G^{(1)}, Q^{(1)}, H^{(1)}\}$ and $S^{(2)} = \{F^{(2)}, G^{(2)}, Q^{(2)}, H^{(2)}\}$, associated with a minimal state-space representation, are observationally equivalent if and only if there is a non-singular $p \times p$ matrix T such that

$$F^{(2)} = TF^{(1)}T^{-1}, H^{(2)} = H^{(1)}T^{-1},\ G^{(2)} = TG^{(1)}, Q^{(2)} = TQ^{(1)}, \tag{5}$$

where T is a matrix of coordinate transformation: $\xi_t^{(2)} = T\xi_t^{(1)}$. The argument for the equivalence conditions (5) is analogous to the proof of Proposition 1 in Wall (1987). The initial conditions are transformed according to the rule:

$$\bar{\xi}_0^{(2)} = T\bar{\xi}_0^{(1)} \text{ and } P_0^{(2)} = TP_0^{(1)}T^{-1}.$$

The non-singular transformation matrix T is uniquely determined for any pair of minimal representations. A minimal state-space representation is uniquely identified if the only admissible transformation is the identity: $T \equiv I_p$.

The set of minimal representations forms a well-defined class of observational equivalence: there are no observationally-equivalent minimal representations of the same model that have identical parameter matrices and different initial states. Specifying a model, which admits an interpretable minimal representation, gives an operational criterion of observational equivalence.

For non-minimal representations, there is no well-defined class of observational equivalence: non-minimal representations, which have identical parameter matrices and different initial conditions, can be observationally equivalent. It means that there are observationally equivalent representations which have the same parameter matrices but different realizations of state variables, induced by different initial conditions. For a non-stationary (integrated) process, which retains the memory of initial conditions, a unique identification of state variables cannot be obtained in a non-minimal representation. However, observationally equivalent structures can be constructed case-by-case using analytical form of distributional moments.

4. NATURAL-RATE MODEL

Consider a semi-structural econometric model of the natural interest rate which is a modification of the Laubach-Williams model admitting an interpretable minimal representation:

Measured Output: $y_t = y_t^* + \tilde{y}_t,$ (6)

Potential Output: $y_t^* = y_{t-1}^* + g_{t-1} + \sigma_{y^*}\varepsilon_{y^*t},$ (7)

IS Equation: $\tilde{y}_t = a_1\tilde{y}_{t-1} + a_2\tilde{y}_{t-2} + a_r(r_{t-1} - r_{t-1}^*) + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t},$ (8)

Phillips Curve: $\Delta\pi_t = b_1\Delta\pi_{t-1} + b_2\Delta\pi_{t-2} + b_3\Delta\pi_{t-3} + b_{\tilde{y}}\tilde{y}_{t-1} + \sigma_\pi\varepsilon_{\pi t},$ (9)

Potential Growth: $g_t = g_{t-1} + \sigma_g\varepsilon_{gt},$ (10)

'Headwinds': $$z_t = \rho_z z_{t-1} + \sigma_z \varepsilon_{zt}, \tag{11}$$

Natural Rate: $$r_t^* = c g_t + z_t, \tag{12}$$

where y_t is the logarithm of the measured output in period t; y_t^* and $\tilde{y}_t$ are the (unobservable) potential output and output gap; g_t is the growth rate of the potential output; π_t is the inflation rate; r_t is the measured real rate of interest; r_t^* is the unobservable natural rate of interest; $(r_t - r_t^*)$ is the real rate gap; z_t is the non-growth component of the natural interest rate ('headwinds'); and ε_{y^*t}, $\varepsilon_{\tilde{y}t}$, $\varepsilon_{\pi t}$, ε_{gt}, ε_{zt} are structural shocks which are independent over time and across variables and have the standard normal distribution. Note that there are no measurement errors in the model – all the shocks enter dynamic equations and drive the dynamics of the system.

The parameters of the model are assumed to satisfy the following restrictions: $a_r < 0$, $b_y > 0$ and $|\rho_z| \leq 1$. The parameters σ_g and σ_z cannot be identified in the model (6)–(12) because of the 'pileup problem', discussed by Stock (1994). Laubach and Williams (2003) estimate these parameters using the median-unbiased estimator described in Stock and Watson (1998). The application of the median-unbiased estimator requires an additional assumption: $\rho_z = 1$. For the estimation of the full system (6)–(12), the parameters σ_g and σ_z are set to be equal to their median-unbiased estimators obtained at preliminary stages.

The potential output y_t^* is modeled as an I(2) variable and the growth rate g_t is assumed to be a random walk. If the persistence parameter ρ_z in equation (11) equals one, the process $\{z_t\}$ is a random walk and, as results from equation (12), the natural rate of interest r_t^* and the growth rate of the potential output g_t can diverge. For values of ρ_z smaller than one, the process $\{z_t\}$ is stationary and the natural rate of interest r_t^* is cointegrated with the growth rate of the potential output g_t.

All the equations in the model (6)–(12), except for the IS equation (8), are equivalent to the equations in the original Laubach-Williams model. The IS equation (8) includes only one lag of the real rate gap $(r_t - r_t^*)$, and this modification guarantees the observability of state variables. The original Laubach-Williams model, including two lags of the real rate gap in the IS equation, fails the observability condition (see Appendix). The failure of the observability condition implies that the natural rate of interest is not uniquely identified in the model. A minimal form of the Laubach-Williams model is derived in the Appendix. However, it is only possible to obtain it by such a transformation of state variables that the Laubach-Williams model loses its original interpretation.

The model (6)–(12) retains the economic interpretation of the Laubach-Williams model and has a minimal state-space representation. It is consistent with all the standard assumptions of the natural rate theory. The potential output y_t^* and its growth rate g_t are exogenous with respect to the output gap $\tilde{y}_t$. The non-growth component of the natural interest rate z_t is also modeled as exoge-

nous with respect to the output gap $\tilde{y}_t$, and, consequently, the natural rate of interest $r_t^* = cg_t + z_t$ is exogenous with respect to the output gap $\tilde{y}_t$. The 'accelerationist' Phillips curve (9) implies that a positive output gap accelerates inflation. Nevertheless, the model can be rewritten in an observationally equivalent form, which admits feedback from the output gap to the non-growth component of the natural interest rate:

Measured Output: $y_t = y_t^* + \tilde{y}_t,$ (6*)

Potential Output: $y_t^* = y_{t-1}^* + g_{t-1} + \sigma_{y^*}\varepsilon_{y^*t},$ (7*)

IS Equation: $\tilde{y}_t = a_1\tilde{y}_{t-1} + \tilde{a}_2\tilde{y}_{t-2} + a_r(r_{t-1} - \tilde{r}_{t-1}^*) + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t},$ (8*)

Phillips Curve: $\Delta\pi_t = b_1\Delta\pi_{t-1} + b_2\Delta\pi_{t-2} + b_3\Delta\pi_{t-3} + b_{\tilde{y}}\tilde{y}_{t-1} + \sigma_\pi\varepsilon_{\pi t},$ (9*)

Potential Growth: $g_t = g_{t-1} + \sigma_g\varepsilon_{gt},$ (10*)

'Headwinds': $\tilde{z}_t = \rho_z\tilde{z}_{t-1} + \alpha(\tilde{y}_{t-1} - \rho_z\tilde{y}_{t-2})\sigma_z\varepsilon_{zt},$ (11*)

Natural Rate: $\tilde{r}_t^* = cg_t + \tilde{z}_t,$ (12*)

where the non-growth component of the natural rate is redefined as $\tilde{z}_t = z_t + \alpha\tilde{y}_{t-1}$ and α is a positive constant. This model can be obtained by an invertible coordinate transformation in a state-space representation of model (6)–(12). The coordinate transformation is a linear transformation of a state vector that generates an observationally equivalent state-space representation, where some or all components of the transformed state vector are different from the components of the original state vector. It also changes structural relations between the components of state vector. The coordinate transformation is achieved by a pre-multiplication of the state vector by an invertible matrix, which determines the changes.

The transformed model retains the same structural form of the IS and the Phillips curve equations. Only one parameter of the IS equation changes: $\tilde{a}_2 = a_2 + \alpha a_r < \mathrm{a}_2$. The modified IS equation (8*) satisfies all the restrictions imposed in original IS equation (8), although it includes a different equilibrium rate: $\tilde{r}_t^* = r_t^* + \alpha\tilde{y}_{t-1}$. The transformed model includes a hysteresis effect: the past demand shocks affect the current value of the equilibrium rate $\tilde{r}_t^*$.

The model (6)–(12) can be written in the state-space form (1)–(2), where

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ \Delta\pi_t \end{bmatrix}, x_t = [r_{t-1}], H = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\xi_t = \begin{bmatrix} y_t^* \\ \tilde{y}_t \\ \tilde{y}_{t-1} \\ g_t \\ z_t \\ \Delta\pi_t \\ \Delta\pi_{t-1} \\ \Delta\pi_{t-2} \end{bmatrix}, F = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_1 & a_2 & -a_r c & -a_r & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho_z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & b_y & 0 & 0 & 0 & b_1 & b_2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, G = \begin{bmatrix} 0 \\ \frac{a_r}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} \sigma_{y^*} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\tilde{y}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_g & 0 & 0 \\ 0 & 0 & 0 & \sigma_z & 0 \\ 0 & 0 & 0 & 0 & \sigma_\pi \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } v_t = \begin{bmatrix} \varepsilon_{y^* t} \\ \varepsilon_{\tilde{y} t} \\ \varepsilon_{g t} \\ \varepsilon_{z t} \\ \varepsilon_{\pi t} \end{bmatrix}.$$

This state-space representation is irreducible (minimal), i.e. it includes no redundant state variables. The irreducibility is implied by the satisfied observability and controllability conditions:

$$rank \begin{bmatrix} H \\ HF \\ HF^2 \\ HF^2 \\ HF^3 \\ HF^4 \\ HF^5 \\ HF^6 \\ HF^7 \end{bmatrix} = 8 \text{ and } rank[Q \;\; FQ \;\; F^2Q \;\; F^3Q \;\; F^4Q \;\; F^5Q \;\; F^6Q \;\; F^7Q] = 8,$$

where the rank is equal to the number of state variables. (The Python routine that implements these rank tests is available from the author upon request).

The vector of state variables ξ_t includes non-stationary (integrated) variables and requires an initialization model (see Hamilton, 1994 or Durbin and Koopman, 2012). The initialization model is not discussed in this paper. For any pair of observationally equivalent minimal state-space representations, there is a unique invertible transformation of initial conditions (which would not be the case for non-minimal representations).

The coordinate transformation matrix, which generates the model (6*)–(12*) from the model (6)–(12), is

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The system matrices of the transformed model $\tilde{H} = HT^{-1}$, $\tilde{G} = TG$ and $\tilde{Q} = TQ$ are identical to the corresponding matrices of the original model, except for the transition matrix $\tilde{F} = TFT^{-1}$:

$$\tilde{F} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_1 & a_2 + \alpha a_r & -a_r c & -a_r & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \alpha & -\alpha\rho_z & 0 & \rho_z & 0 & 0 & 0 \\ 0 & b_{\tilde{y}} & 0 & 0 & 0 & b_1 & b_2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The observational equivalence follows from the minimality of the original state-space representation and the invertibility of transformation matrix T.

For a non-minimal model, such as the original Laubach-Williams model, there are multiple realizations of state variables (including the natural rate r_t^*) that are induced by different initial conditions and are observationally equivalent for a given structure (see Appendix).

5. CONCLUSIONS

This paper demonstrates that it is possible to transform a structural model of the natural (equilibrium) interest rate in such a way as to obtain an interpretable observationally equivalent model in which the redefined 'unnatural' interest rate is different, because it depends on past output gaps.

Specifying a model that admits a minimal state-space representation allows a class of observationally equivalent models to be defined. For a well-defined class of observationally equivalent models it is easier to identify interpretable alternatives and envision model modifications which would exclude such alternatives.

The cause of the model non-uniqueness is the inherent unobservability of the natural (equilibrium) rate, which is determined by other unobservable variables and does not directly enter into an equation for an observable variable. It allows redefining the equilibrium rate by reshuffling other unobservable variables without losing information about observable variables. A potential solution to this

problem is to augment the information set which is used for the identification and estimation of the natural rate, for example, the dynamics of the non-growth component of the natural rate (z_t) can be explained by some observable variables. Yellen (2015) lists some of the 'headwinds' that determine the non-growth component of the natural interest rate. Augmenting a natural-rate model with exogenous observable variables that explain the dynamics of z_t would restrict the class of observationally equivalent models.

The unobservability problem in the Laubach-Williams model can potentially be solved by imposing economically-motivated restrictions on the initial values of state variables, or by re-specifying the dynamics of natural-rate components. For example, specifying g_t and z_t as second-order autoregressive processes may solve the problem of unobservability. However, it would require either the estimation or the calibration of additional parameters.

The issue of the model non-uniqueness becomes important when there is an observationally equivalent model which admits a meaningful alternative interpretation. In the example presented in this paper, the hysteresis effect in the transformed model can explain the persistent shift in the level of the equilibrium interest rate caused by a demand-driven recession. This interpretation is consistent with persistently low real interest rates in many advanced industrial economies in the aftermath of the financial crisis of 2007–2008. The model, in which there is a feedback from the output gap to the equilibrium interest rate, has particular implications for the monetary policy, namely, it calls for a more active monetary policy response to contractionary demand shocks.

APPENDIX. MINIMAL REPRESENTATION OF THE LAUBACH-WILLIAMS MODEL

Consider the original Laubach-Williams model:

Measured Output: $$y_t = y_t^* + \tilde{y}_t, \tag{13}$$

Potential Output: $$y_t^* = y_{t-1}^* + g_{t-1} + \sigma_{y^*}\varepsilon_{y^*t}, \tag{14}$$

IS Equation: $$\tilde{y}_t = a_1\tilde{y}_{t-1} + a_2\tilde{y}_{t-2} + \frac{a_r}{2}\sum_{j=1}^{2}\left(r_{t-j} - r_{t-j}^*\right) + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t}, \tag{15}$$

Phillips Curve: $$\pi_t = b_\pi\pi_{t-1} + (1-b_\pi)\bar{\pi}_{t-2,4} + b_{\tilde{y}}\tilde{y}_{t-1} + \sigma_\pi\varepsilon_{\pi t}, \tag{16}$$

Potential Growth: $$g_t = g_{t-1} + \sigma_g\varepsilon_{gt}, \tag{17}$$

'Headwinds': $$z_t = z_{t-1} + \sigma_z\varepsilon_{zt}, \tag{18}$$

Natural Rate: $$r_t^* = cg_t + z_t, \tag{19}$$

where y_t is the logarithm of the measured output in the period *t*; y_t^* and $\tilde{y}_t$ are the (unobservable) potential output and output gap; g_t is the growth rate of the potential output; π_t is the inflation rate; r_t is the measured real rate of interest; $\bar{\pi}_{t-2,4}$ is the average of inflation over periods $t-2$, $t-3$ and $t-4$; $\bar{\pi}_{t-2,4} = (\pi_{t-2} + \pi_{t-3} + \pi_{t-4})/3$; r_t^* is the unobservable natural rate of interest; $(r_t - r_t^*)$ is the real rate gap; z_t is the non-growth component of the natural interest rate ('headwinds'); and ε_{y^*t}, $\varepsilon_{\tilde{y}t}$, $\varepsilon_{\pi t}$, ε_{gt}, ε_{zt} are structural shocks which are independent over time and across variables and have the standard normal distribution.

Using the definition of $\bar{\pi}_{t-2,4}$ and the restriction on inflation lags in the equation (16), it can be rewritten as:

$$\textit{Phillips Curve:} \qquad \Delta\pi_t = b_1\Delta\pi_{t-1} + b_2\Delta\pi_{t-2} + b_3\Delta\pi_{t-3} + b_{\tilde{y}}\tilde{y}_{t-1} + \sigma_\pi\varepsilon_{\pi t}, \qquad (16^*)$$

where $b_1 = (b_\pi - 1)$, $b_2 = \frac{2}{3}(b_\pi - 1)$ and $b_3 = \frac{1}{3}(b_\pi - 1)$. The equations (16) and (16*) are equivalent representations of the Phillips curve.

Since all disturbances (ε_{y^*t}, $\varepsilon_{\tilde{y}t}$, $\varepsilon_{\pi t}$, ε_{gt} and ε_{zt}) in the model (13)–(19) are structural shocks entering dynamic equations, it can be written in the state-space representation (1)–(2) where

$$\xi_t = \begin{bmatrix} y_t^* \\ \tilde{y}_t \\ \tilde{y}_{t-1} \\ g_t \\ g_{t-1} \\ z_t \\ z_{t-1} \\ \Delta\pi_t \\ \Delta\pi_{t-1} \\ \Delta\pi_{t-2} \end{bmatrix}, F = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_1 & a_2 & -\frac{a_r c}{2} & -\frac{a_r c}{2} & -\frac{a_r}{2} & -\frac{a_r}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho_z & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & b_y & 0 & 0 & 0 & 0 & 0 & b_1 & b_2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & 0 \\ \frac{a_r}{2} & \frac{a_r}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ \Delta\pi_t \end{bmatrix}, x_t = \begin{bmatrix} r_{t-1} \\ r_{t-2} \end{bmatrix}, H = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} \sigma_{y^*} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\tilde{y}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_g & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_z & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\pi \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } v_t = \begin{bmatrix} \varepsilon_{y^*t} \\ \varepsilon_{\tilde{y}t} \\ \varepsilon_{gt} \\ \varepsilon_{zt} \\ \varepsilon_{\pi t} \end{bmatrix}.$$

This state-space representation is non-minimal: it includes two redundant (unobservable) states. The redundancy follows from the failure of the observability rank condition (The Python routine that implements the rank test is available from the author upon request):

$$rank \begin{bmatrix} H \\ HF \\ HF^2 \\ HF^2 \\ HF^3 \\ HF^4 \\ HF^5 \\ HF^6 \\ HF^7 \\ HF^8 \\ HF^9 \end{bmatrix} = 8 < 10.$$

The source of the redundancy is the imbalance in the dynamics of state variables: although variables g_t and z_t are defined as first-order autoregressive processes (random walks), two lags of each variable are included in the IS equation:

$$\tilde{y}_t = a_1\tilde{y}_{t-1} + a_2\tilde{y}_{t-2} + \frac{a_r}{2}\sum_{j=1}^{2}\left(r_{t-j} - r^*_{t-j}\right) + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t} =$$

$$a_1\tilde{y}_{t-1} + a_2\tilde{y}_{t-2} + \frac{a_r}{2}\sum_{j=1}^{2}\left(r_{t-j} - cg_{t-j} - z_{t-j}\right) + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t}.$$

A minimal representation can be obtained by applying a decomposition which is analogous to the decomposition used in Youla (1966, Corollary 2). The coor-

dinate transformation of the state vector, which can be implemented to obtain the decomposition, is given by the invertible matrix below:

$$T = \left[\frac{T^{(0)}}{T^{(1)}}\right] = \left[\begin{array}{cccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \frac{a_r c}{2a_2} & 0 & \frac{a_r}{2a_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\rho_z} & -1 & 0 & 0 & 0 \end{array}\right].$$

The sub-matrix $T^{(0)}$ composed of the first eight rows of the matrix T generates a minimal state vector $\xi_t^{(0)} = T^{(0)}\xi_t$. The sub-matrix $T^{(1)}$ composed of the last two rows of the matrix T generates a vector of redundant states $\xi_t^{(1)} = T^{(1)}\xi_t$ which affect neither the minimal state vector $\xi_t^{(0)}$ nor the vector of explained variables y_t. Because of the redundancy, the transformation matrix T is not unique.

The structure of the minimal state-space system is

$$\xi_t^{(0)} = \begin{bmatrix} y_t^* \\ \tilde{y}_t \\ w_t \\ g_t \\ z_t \\ \Delta\pi_t \\ \Delta\pi_{t-1} \\ \Delta\pi_{t-2} \end{bmatrix}, F^{(0)} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & a_1 & a_2 & -\frac{a_r c}{2} & -\frac{a_r}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{a_r c}{2a_2} & \frac{a_r}{2a_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho_z & 0 & 0 & 0 \\ 0 & b_y & 0 & 0 & 0 & b_1 & b_2 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, G^{(0)} = \begin{bmatrix} 0 & 0 \\ \frac{a_r}{2} & \frac{a_r}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ \Delta\pi_t \end{bmatrix}, x_t = \begin{bmatrix} r_{t-1} \\ r_{t-2} \end{bmatrix}, H = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

where neither the new state variable $w_t = \tilde{y}_{t-1} + \frac{a_r}{2a_2}(cg_{t-1} + z_{t-1})$ nor the transformed equation describing the dynamics of the output gap have a meaningful economic interpretation:

$$\tilde{y}_t = a_1\tilde{y}_{t-1} + a_2 w_{t-1} + \frac{a_r}{2}(r_{t-1} - r_{t-1}^*) + \frac{a_r}{2}\, r_{t-2} + \sigma_{\tilde{y}}\varepsilon_{\tilde{y}t}. \qquad (20)$$

The equation (20) cannot be interpreted as an IS equation. If we try to restore the original model (13)–(19) from its minimal form by inverting the transformation T, then any choice of initial values y_{-1}, g_{-1} and z_{-1} such that $w_0 = \tilde{y}_{-1} + \frac{a_r}{2a_2}(cg_{-1} + z_{-1})$ does not change will generate an observationally equivalent model with a different natural rate of interest.

Hendry (1995, p. 36) defines both the uniqueness and the interpretability as necessary conditions for the model identification. In the original Laubach-Williams model, state variables are consistent with the assumed interpretation, but they are not unique (multiple realizations of the natural interest rate are possible in the same model). If the Laubach-Williams model is reduced to a minimal representation, state variables are uniquely identified, but the model loses the assumed interpretation. The problem of identification should be reconsidered at an earlier stage, when the model specification is selected. The model specification (6)–(12) guarantees both the uniqueness of state variables and their interpretability.

REFERENCES

Beyer, R. C. M., Wieland V. (2017), Instability, imprecision and inconsistent use of the equilibrium real interest rate estimates*, CEPR Discussion Papers,* DP11927.

Constancio, V. (2016), The challenge of low real interest rates for monetary policy, Lecture at the Macroeconomic Symposium at Utrecht School of Economics, https://www.ecb.europa.eu/press/key/date/2016/html/sp160615.en.html.

Durbin J., Koopman S. J., (2012), *Time Series Analysis by State-Space Methods: Second Edition*, Oxford University Press, Oxford.

Fiorentini, G., Galesi, A., Pérez-Quirós, G., Sentana E. (2018), The rise and fall of the natural interest rate, *Banco de España Working Papers,* 1822.

Fries, S., Mésonnier, J.-S., Mouabbi, S., Renne, J.-P. (2018), National natural rates of interest and the single monetary policy in the euro area, *Journal of Applied Econometrics*, 33(6), 763–769.

Garnier, J., Wilhelmsen, B.-R. (2009), The natural rate of interest and the output gap in the Euro Area: a joint estimation, *Empirical Economics*, 36, 297–319.

Grossman, V., Martínez-García, E., Wynne, M., Zhang, R. (2019), Ties that bind: estimating the natural rate of interest for small open economies, *Globalization and Monetary Policy Institute Working Papers,* 359.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton NJ.

Hamilton, J. D, Harris, E. S., Hatzius, J., West, K. D. (2016), The equilibrium real funds rate: past, present and future, *IMF Economic Review*, 64(4), 660–707.

Hendry, D. F. (1995), *Dynamic Econometrics*, Oxford University Press, Oxford.

Holston, K., Laubach, T., Williams, J. C. (2017), Measuring the natural rate of interest: international trends and determinants, *Journal of International Economics*, 108(S1), 59–75.

Kiley, M. T. (2019), The global equilibrium real interest rate: concepts, estimates, and challenges, *Finance and Economics Discussion Series,* 2019-076, Board of Governors of the Federal Reserve System, Washington.

Laubach, T., Williams, J. C. (2003), Measuring the natural rate of interest, *Review of Economics and Statistics*, 85(4), 1063–1070.

Laubach, T., Williams, J. C. (2016.), Measuring the natural rate of interest redux, *Business Economics*, 51(2), 57–67.

Lewis, K. F., Vazquez-Grande, F. (2017), Measuring the natural rate of interest: alternative specifications, *Finance and Economic Discussion Series,* 2017–059, Board of Governors of the Federal Reserve System, Washington.

Mésonnier, J.-S., Renne, J.-P. (2007), A time-varying 'natural' rate of interest for the euro area, *European Economic Review*, 51, 1768–1784.

Preston, A. J. (1978), Concepts of structure and model identifiability for econometric systems, [in] Bergstrom, A. R., Catt, A. J. L., Peston, M. H., Silverstone, B. D. J. (eds.), *Stability and Inflation: A Volume of Essays to Honour the Memory of A.W.H. Phillips*, 275–97, Wiley, New York.

Rothenberg, T. J. (1971), Identification in Parametric Models, *Econometrica*, 39(3), 577–591.

Sargent, T. J. (1976), The observational equivalence of natural and unnatural rate theories of macroeconomics, *Journal of Political Economy*, 84(3), 631–640.

Stock, J. H. (1994), Unit roots, structural breaks and trends, [in] R.F. Engle, R. F., McFadden, D. L. (eds.), *Handbook of Econometrics*, 4, Elsevier Science B.V., 2739–2841.

Stock, J. H., Watson, M. W. (1998), Median unbiased estimation of coefficient variance in a time-varying parameter model, *Journal of American Statistical Association*, 93(441), 349–357.

Taylor, J. B., Wieland, V. (2016), Finding the equilibrium real interest rate in a fog of policy deviations, *Business Economics*, 51(3), 147–154.

Wall, K. D. (1987), Identification theory for varying coefficient regression models, *Journal of Time Series Analysis*, 8(3), 359–371.

Wicksell, K. (1898), *Interest and Prices*, Sentry Press, New York.

Williams, J. C. (2003), The natural rate of interest, *FRBSF Economic Letters,* 2003–32.

Wynne, M. A., Zhang, R. (2018a), Estimating the natural rate of interest in an open economy, *Empirical Economics* 55(3), 1291–1318.

Wynne, M. A. and Zhang, R. (2018b), Measuring the world natural rate of interest, *Economic Inquiry* 56(1), 530–544.

Yellen, J. (2015), Normalizing Monetary Policy: Prospects and Perspectives, Speech at the 'New Normal Monetary Policy', Conference, https://www.federalreserve.gov/newsevents/speech/yellen20150327a.htm.

Youla, D. C. (1966), The synthesis of linear dynamical systems from prescribed weighting patterns, *SIAM Journal of Applied Mathematics*, 14(3), 527–549.

DOI: 10.5604/01.3001.0013.9129

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Grażyna DEHNEL[1]
Łukasz WAWROWSKI[2]

Estimation of the average wage in Polish small companies using the robust approach

Abstract. *There is a growing demand for multivariate economic statistics for cross-classified domains. In business statistics, this demand poses a particular challenge given the specific character of the population of enterprises, which necessitates searching for methods of analysis that would represent the robust approach to estimation, where auxiliary variables could be utilised. The adoption of new solutions in this area is expected to increase the scope of statistical output and improve the precision of estimates. The study presented in the paper furthers this goal, as it is focused on testing the application of a robust version of the Fay-Herriot model, which makes it possible to meet the assumption of normality of random effects under the presence of outliers. These alternative models are supplied to estimate the parameters of small firms operating in 2012. Variables from administrative registers were used as auxiliary variables, which made the estimation process more comprehensive. The paper refers to small area estimation methods. The variables of interest are estimated at a low level of aggregation represented by the cross-section province and NACE sections.*

Keywords: robust estimation, business statistics, small area estimation, Fay-Herriot model

JEL Classification: C40, C13, C40, C51, M20

1. INTRODUCTION

One of the conditions for economic growth is the development of entrepreneurship. Nowadays, however, in the fast-changing social, economic and legal environment, it is not easy to do business: customer needs keep changing, different markets are undergoing integration and business environment is getting increasingly competitive. To meet these challenges, entrepreneurs have to interact with various actors and exchange information; they need access to detailed information at a low level of aggregation, which enables them to react quickly to market changes.

[1] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, al. Niepodległości 10, 61-875, Poznan, Poland, corresponding author – e-mail: g.dehnel@ue.poznan.pl, ORCID: https://orcid.org/0000-0002-0072-9681.

[2] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, al. Niepodległości 10, 61–875, Poznan, Poland, ORCID: https://orcid.org/0000-0002-1201-5344.

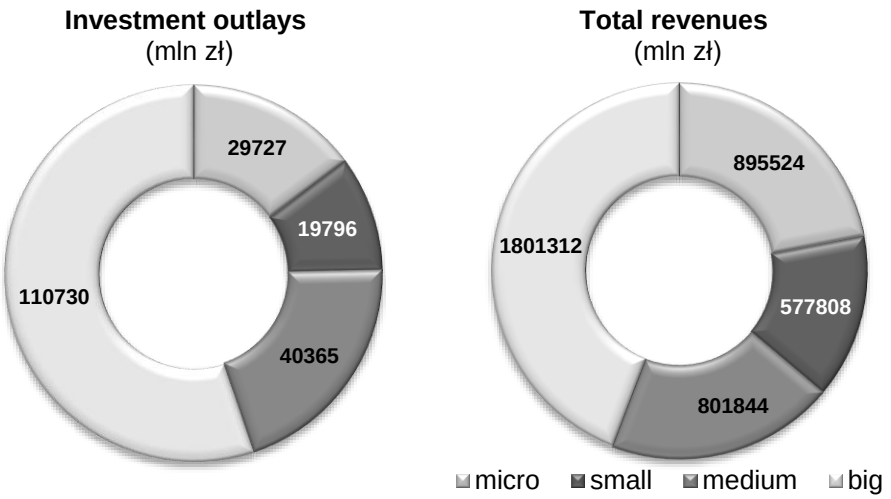
The growing demand for information for small domains has called for new estimation methods that would satisfy consumers' needs in this regard. In the case of economic statistics, the estimation of key variables proves particularly challenging due to problems such as strong asymmetry and high variation and concentration, which make it difficult to retain the properties of classical estimators used in sample surveys. To overcome these problems, there have been attempts to apply robust indirect estimation techniques using auxiliary variables from additional data sources, which could yield more reliable estimates than those obtained by means of direct estimation. This paper contributes to this approach, as its aim is to test the usefulness of the application of one of the methods from the realm of small area statistics to the estimation of the average salary in the enterprise sector according to province and NACE[3] section, utilising information collected in administrative registers.

The paper consists of four parts. The first part is devoted to the characteristics of the Polish small business. The second part describes data sources used for the estimation and provides details of the empirical study. The third part applies methodological considerations to the analysis. The fourth part summarizes the results of the study and presents their interpretation. The study focuses on small enterprises employing from 10 to 49 persons. Its aim is to estimate the average wage in these companies using a robust version of the Fay-Herriot model and auxiliary variables from administrative registers (Fay and Herriot, 1979; Sinha and Rao, 2009). The study is the continuation of the antecedent research on this subject by Dehnel and Wawrowski (2018).

The structure of the Polish business sector has remained stable for many years, where small enterprises have constituted less than 3 percent of the entire sector. Nevertheless, they have played a significant, and, in some respects, a crucial role in the economy. It is because small firms, which are free from corporate connections and dependencies, are able to compete with the largest units. They are legally and economically independent to a considerable extent, and also relatively flexible thanks to tight cost control, quick responsiveness to changing market requirements, and the ability to quickly implement innovations. In 2015, small companies invested almost PLN 20 billion (9.9% of the total value of investments in the enterprise sector), cf. Fig. 1. They acted according to their own strategies, strove to achieve their own goals, often taking financial risk. Their revenues accounted for about a quarter of the revenue of the entire small and medium enterprises sector (SME sector).

[3] NACE – The Statistical Classification of Economic Activities in the European Community

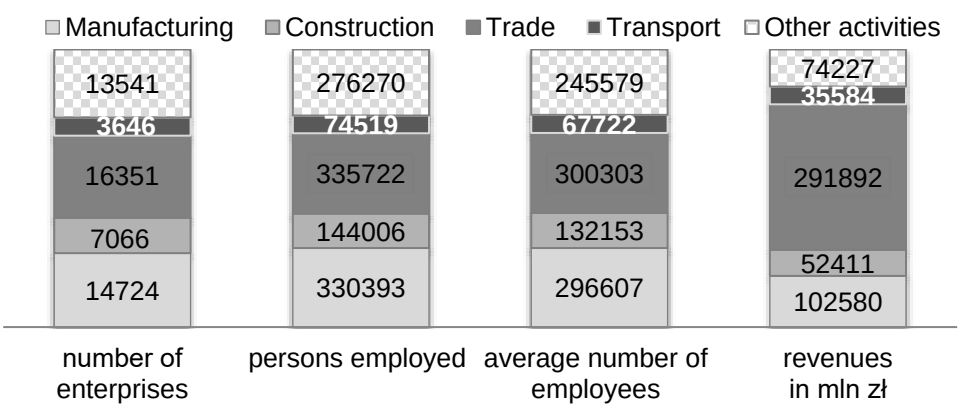
Figure 1. Enterprises' characteristics by size class as of 31 Dec 2015 (millions of PLN)



S o u r c e: based on "Statistics Poland's" study (GUS, 2017).

From the point of view of business classification, the most important sections in this sector are: manufacturing, construction, wholesale and retail trade (trade), and transport and storage (transport). These sections account for over 75% of all small businesses, produce almost 90% of the total revenue of the sector cf. Fig. 2. and also provide 86% of all the jobs in the small business sector (GUS, 2017).

Figure 2. Small enterprises' characteristics by NACE section as of 31 Dec 2015



S o u r c e: based on "Statistics Poland's" study (GUS, 2017).

2. DESCRIPTION OF THE STUDY

Data for the present analysis has been drawn from the DG1[4] survey carried out by the Statistical Office in Poznań. The survey is conducted in the form of reports that all large and medium-size enterprises as well as a 10-percent sample of small companies have to submit every month, and whose objective is to collect updated basic indicators of the economic activity.

For the purpose of the study, the scope of data collected from the DG1[4] survey was limited to the statistics of small enterprises operating in August 2012 – the period determined by the availability of data. The average wage was the target variable, while net revenues in 2011 taken from the Ministry of Finance's register and the number of enterprises per 10,000 population in 2011 taken from the REGON register were the auxiliary variables.

The data concerning the average wage in small companies from the manufacturing, construction, trade and transportation sections, published by Statistics Poland, is available only at the country level. For this reason, as well as being aware of the growing demand for more detailed information voiced by data users, the authors decided to carry out a study whose goal was to estimate certain variables at the level of province (NUTS 2), thus the target domain for estimation in the paper is a province cross-classified by NACE section (Dehnel, 2017).

3. ROBUST FAY-HERRIOT MODEL

The Fay-Herriot model belongs to a class of area-level models, which means that it utilises aggregated data instead of unit-level information. This approach was developed in 1979 as a tool for estimating income for small areas in the United States (Fay and Herriot, 1979). The Fay-Herriot model is constructed in two stages. Firstly, it is assumed that the direct estimator is unbiased and can be written as the sum of the true value of the estimated parameter and the random error:

$$\hat{\theta}_d = \theta_d + e_d, \tag{1}$$

where $e_d \overset{iid}{\sim} N(0, \sigma_{ed}^2)$. In practice, the variance σ_{ed}^2 is unknown and has to be estimated on the basis of the survey data. The direct estimator used most frequently in the Fay-Herriot model is the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952), which has also been used in this study.

In the second stage, the true value of the parameter is treated as a dependent variable in the linear model with an area random effect:

[4] DG1 – the largest survey in Polish short-term business statistics. It collects data from businesses employing over 9 people.

$$\theta_d = x_d^T \beta + u_d, \tag{2}$$

where x_d is a vector of auxiliary information for area d, β is a vector of regression parameters and u_d is an area random effect with the distribution $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$.

By combing equations (1) and (2), we obtain the formula of the Fay-Herriot model:

$$\hat{\theta}_d = {x_d}^T \beta + u_d + e_d. \tag{3}$$

The estimator of the Fay-Herriot model is known as EBLUP (Empirical Best Linear Unbiased Predictor) and is expressed by the following formula:

$$\hat{\theta}_d^{FH} = x_d^T \hat{\beta} + \hat{u}_d = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) x_d^T \hat{\beta}, \; d = 1, \ldots, D \tag{4}$$

where

$$\hat{\beta} = \left(\sum_{d=1}^D \hat{\gamma}_d x_d x_d^T\right)^{-1} \sum_{d=1}^D \hat{\gamma}_d x_d \hat{\theta}_d \text{and } \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{ed}^2}. \tag{5}$$

EBLUP is a weighted average of the direct estimator and the regression-synthetic estimator. The weight $\hat{\gamma}_d$ measures the uncertainty of the regression component. If the sample variance estimator $\hat{\sigma}_{ed}^2$ is small, then a larger part of the final estimate will be contributed by the direct estimator $\hat{\theta}_d$ (Boonstra and Buelens, 2011). The between-area variance $\hat{\sigma}_u^2$ as a sample variance is also unknown and has to be estimated, which can be done using many techniques, e.g. the Fay-Herriot method, Prasad-Rao method, ML or REML described in Chapter 6 of Rao book (2014, p. 126–129).

The robust version of the Fay-Herriot model uses the Huber (1981) influence function to restrict the influence of u_d and e_d.

Let us replace the estimates of $\hat{\sigma}_{ed}^2$ and $\hat{\sigma}_u^2$ with covariance matrices Σ_e and Σ_u and let $V = \Sigma_e + \Sigma_u$. Then the vector of the fixed effects β is expressed by:

$$\beta = (X^T V^{-1} X)^{-1} X V^{-1} y \tag{6}$$

and random effects vector u is:

$$u = \Sigma_u Z^T V^{-1} (y - X\beta). \tag{7}$$

It is demonstrated that equations (6) and (7) can be transformed into:

$$X^T V^{-1} (y - X\beta) = 0 \tag{8}$$

and

$$\Sigma_u Z^T V - 1(y - X\beta) - u = 0. \tag{9}$$

Sinha and Rao (2009) proposed a robust version of equations (8) and (9):

$$X^T V^{-1} U^{\frac{1}{2}} \psi(U^{\frac{1}{2}}(y - X\beta)) = 0, \tag{10}$$

where $U = diag(V)$. A robust random effects vector is defined by:

$$\begin{gathered}\psi((y - X\beta)^T U^{\frac{1}{2}}) U^{\frac{1}{2}} V^{-1} (\partial V / \partial\theta) V^{-1} U^{\frac{1}{2}} \psi(U^{\frac{1}{2}}(y - X\beta)) \\ = \mathrm{tr}(\mathrm{D}^{\psi}(\partial \mathrm{V} / \partial \theta)),\end{gathered} \tag{11}$$

where $\partial V/\partial\theta$ is the first order partial derivative of V with respect to the variance component θ and for $Z \sim N(0,1)$, $D^{\psi} = E(\psi^2(Z))V^{-1}$.

Moreover, Warnholz (2016) proposed a modification of the above equation in which only diagonal elements of V matrix are used to standardise the residuals. In the robust Fay-Herriot model this matrix is diagonal, but the transformation can be useful in models with correlated random effects, e. g. SAR(1) and AR(1), where calculations are likely to be time-consuming.

Robust EBLUP is expressed by the formula:

$$\hat{\theta}_d^{RFH} = x_d^T \hat{\beta}^{\psi} + \hat{u}_d^{\psi}, \; d = 1, \ldots, D. \tag{12}$$

For unsampled domains, and where the between-area variance equals zero, the indirect estimation relies only on the regression component.

To estimate the mean square error (MSE) for Fay-Herriot model, we can use the parametric bootstrap method proposed by González-Manteiga et al. (2008). The algorithm proceeds along the following steps:

1. fit the model to obtain estimates of $\hat{\sigma}_u^2$ and $\hat{\beta}$;
2. generate a vector of u^* with $N(0, \hat{\sigma}_u^2)$ and calculate $\theta^* = X\hat{\beta} + u^*$;
3. generate a vector of e^* with $N(0, \hat{\sigma}_{ed}^2)$;
4. construct a bootstrap data vector of $\hat{\theta}^* = \theta^* + e^* = X\hat{\beta} + u^* + e^*$;
5. fit the model to bootstrap data $\hat{\theta}^*$ to obtain new estimates of $\hat{\sigma}_u^{2*}$ and $\hat{\beta}^*$;
6. calculate $\hat{\theta}^{*B}$ taking into account values obtained in step 5;
7. repeat steps 2-6 B times, assuming that $\theta^{*(b)}$ is the true value, and $\hat{\theta}^{*(b)}$ are EBLUP estimates obtained in b-th bootstrap replication;
8. the MSE estimator of $\hat{\theta}$ is expressed by:

$$MSE(\hat{\theta}) = B^{-1}\sum_{b=1}^{B}[\hat{\theta}^{*(b)} - \theta^{*(b)}]^2. \tag{13}$$

In the case of the robust Fay-Herriot, model parameter estimates are replaced by their robust versions $\hat{\beta}^{\psi}$, $\hat{\sigma}_u^{2\,\psi}$ and $\hat{\sigma}_{ed}^{2\,\psi}$ and the Robust Fay-Herriot model is calculated in step 5 of the above algorithm (Sinha and Rao, 2009).

Given the MSE, one can calculate relative root mean square error, which is a common measure of precision used in all approaches:

$$RRMSE\left(\hat{\theta}\right) = \frac{\sqrt{MSE\left(\hat{\theta}\right)}}{\hat{\theta}}. \tag{14}$$

4. ESTIMATION RESULTS AND ASSESSMENT OF THEIR PRECISION

Out of the total of 21 NACE sections, the following four were selected: manufacturing, construction, trade and transportation − as this particular combination yielded samples of different sizes. Tabl. 1 presents descriptive statistics of the sample size in the selected sections.

TABLE 1. SAMPLE SIZE BY NACE SECTION

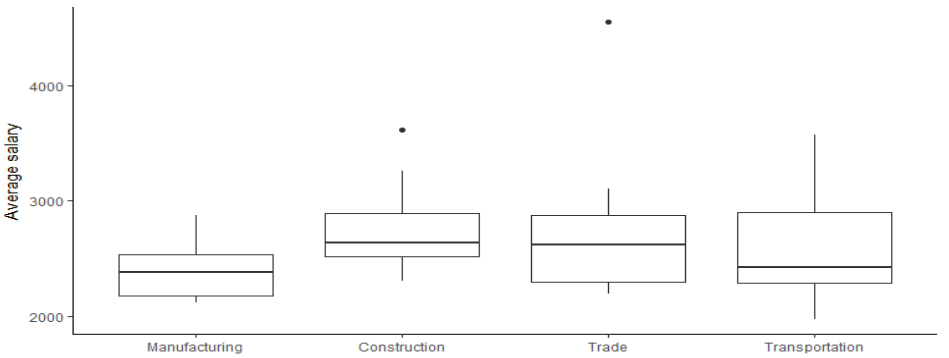
NACE section	Minimum	Median	Mean	Maximum
Manufacturing	129	222	245	440
Construction	41	77	93	197
Trade	131	216	256	562
Transportation	19	31	40	101

S o u r c e: based on data from the DG1 survey.

As the figures demonstrate, the biggest samples were selected for the trade section. The largest of them consisted of 562 enterprises and came from the Śląskie province, whereas the second largest, of 547 companies, came from Mazowieckie. Within the manufacturing section, the Wielkopolskie province provided the biggest sample, of 440 enterprises. The smallest sample of all, which consisted of 19 enterprises, was selected for the transportation section in the Opolskie province. The smallest sample for the construction section consisted of 41 enterprises and was drawn from the Podlaskie province.

The first step in the analysis was to produce direct estimates of the variable of interest for all target domains i.e. province and 4 sections. Fig. 3 presents the distribution of the estimates.

Figure 3. Distribution of the average wage estimates by NACE section



S o u r c e: based on data from the DG1 survey and the administrative register.

Fig. 3 shows two province outliers. In both cases, it is the Mazowieckie province, with the average wage at 3614 PLN for the construction section and 4550 PLN for the trade section. The smallest value of the average wage for the construction section is 2306 PLN, in the Świętokrzyskie province, whereas the smallest value for the trade section is 2194 PLN, in the Lubelskie province. The smallest value of the average wage, 1970 PLN, was estimated for the Podkarpackie for the transportation section, whereas the largest value, in this section, – 3570 PLN, was estimated for the Pomorskie province. The manufacturing section was characterized by the smallest range of wages of all the four sections. The smallest value of the average wage in this section was 2114 PLN (in the Podlaskie province), and the largest amounted to 2875 PLN (in the Mazowieckie). The figures indicate that the average wages in all but one section assume the largest values in the Mazowieckie province. A similar level of the median values of the average wage – 2637 PLN and 2622 PLN – were observed in the construction and trade sections, respectively, as well as in the manufacturing and transportation sections, where they reached 2378 PLN and 2421 PLN, respectively.

In addition to the distribution of direct estimates, it is important to analyse the precision of these estimates. Variances of direct estimates were calculated using the bootstrap method implemented in the survey R package (Lumley 2004). Tabl. 2 presents descriptive statistics of relative root mean square errors (RRMSE) of the direct estimates of the average wage.

TABLE 2. DESCRIPTIVE STATISTICS OF RRMSE OF ESTIMATES BY NACE SECTION (IN %)

NACE section	Minimum	Median	Mean	Maximum
Manufacturing	2.2	3.0	3.0	3.9
Construction	2.9	5.4	5.4	7.2
Trade	2.9	3.4	3.5	4.3
Transportation	5.0	8.7	9.9	21.4

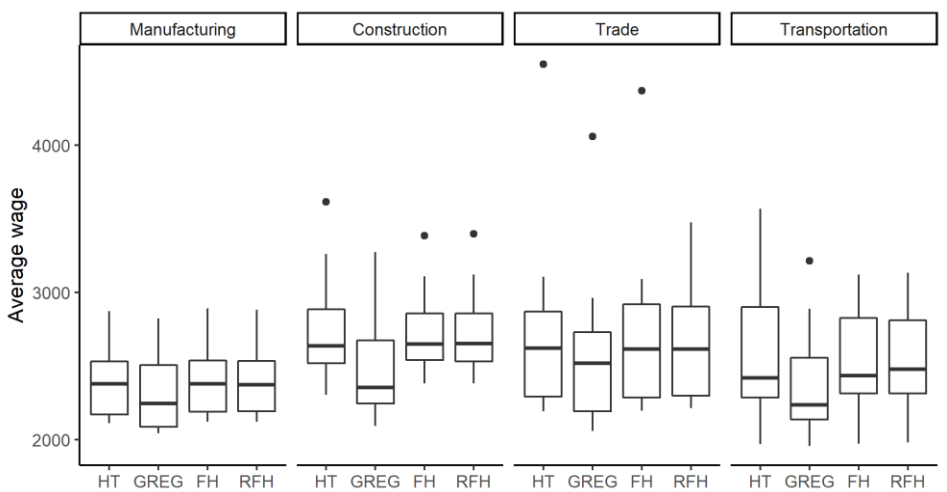
S o u r c e: based on data from the DG1 survey and the administrative register.

Direct estimates of the average wage in all the sections except transportation are relatively precise. The maximum value of the RRMSE for these three sections does not exceed 7.2% (Warmińsko-Mazurskie province). In the case of the transportation section, however, the maximum relative root mean square error amounts to over 20%. This particular value was observed in the Opolskie, where the sample was of the smallest size. According to Statistics Poland's guidelines, estimates can only be published if their RRMSE falls below 10% for planned domains (GUS, 2013; Eurostat, 2013).

To obtain more precise estimates, the authors applied indirect methods of estimation – the Fay-Herriot model (FH) and the robust Fay-Herriot model (RFH). In the modelling process, they used data concerning the average wage in 2011 from the registers of the Ministry of Finance, and concerning the number of enterprises per 10,000 population from the REGON register.

The distributions of the estimates based on the direct estimator (HT – Horvitz and Thompson, 1952), GREG (Dehnel, 2017), the Fay-Herriot model (Dehnel et al., 2017) and Robust Fay-Herriot model are presented in Fig. 4.

Figure 4. Distribution of the average wage estimates by NACE section and estimator



S o u r c e: based on the data from the DG1 survey and the administrative register.

For all the four approaches, the distribution of estimates is similar. The most visible change in the distribution can be observed for the maximum value of the average wage in the trade section. The Horvitz-Thompson estimate was 4550 PLN, the value estimated by the Fay-Herriot model 4367 PLN, and by the robust Fay-Herriot – 3476 PLN.

The precision of estimates can be assessed on the basis of the values of relative root mean square errors presented in Tabl. 3, listed for each section and estimator.

TABLE 3. DESCRIPTIVE STATISTICS OF RRMSE OF ESTIMATES BY NACE SECTION AND ESTIMATOR (IN %)

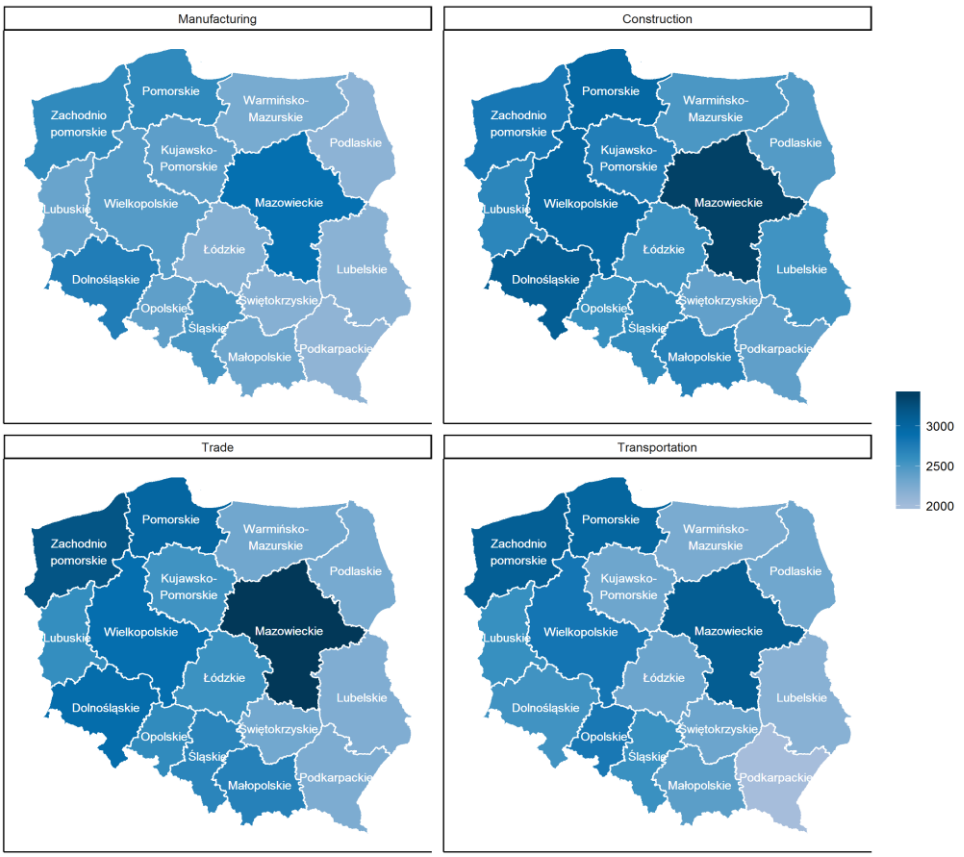
NACE section	Estimator	Minimum	Median	Mean	Maximum
Manufacturing	HT	2.2	3.0	3.0	3.9
Manufacturing	GREG	1.9	2.5	2.6	3.5
Manufacturing	FH	2.0	2.6	2.6	3.2
Manufacturing	RFH	2.0	2.6	2.6	3.2
Construction	HT	2.9	5.4	5.4	7.2
Construction	GREG	2.8	4.8	5.0	7.3
Construction	FH	2.8	4.6	4.6	5.8
Construction	RFH	2.8	4.6	4.6	5.7
Trade	HT	2.9	3.4	3.5	4.3
Trade	GREG	2.4	3.0	3.0	3.6
Trade	FH	2.8	3.3	3.4	4.4
Trade	RFH	2.4	2.8	2.8	3.5
Transportation	HT	5.0	8.7	9.9	21.4
Transportation	GREG	4.8	7.4	8.5	21.3
Transportation	FH	4.1	6.1	6.1	8.6
Transportation	RFH	3.5	5.8	6.0	10.0

S o u r c e: based on data from the DG1 survey and the administrative register.

The application of indirect methods of estimation made it possible to reduce the RRMSE of the average wage for unplanned domains, i.e. provinces cross-classified with NACE sections. The RRMSE of the estimates obtained using the Fay-Herriot model are always lower than the precision of direct estimates. Robust Fay-Herriot estimates for all sections are, on average, either equally or more precise than those based on the Fay-Herriot model. The exception here is the transportation section, where the maximum RRMSE value is higher than that estimated by the Fay-Herriot model. This has been caused by a small sample from the Opolskie province. In general, none of the estimates exceed the 10% threshold set by Statistics Poland.

It is worth mentioning, though, that the MSE estimators are also biased, but this aspect is not analysed in detail in literature on small area estimation (Krzciuk, 2017). The size of the error can be estimated using the Monte Carlo simulation, but to do this, one would have to know the value of the estimated quantity for the whole population (Żądło, 2008, Żądło, 2012). Such information was not available for this study. Another step in the assessment of the obtained results is the analysis of spatial variation. Fig. 5 visualises the average wage across provinces for the four NACE sections.

Figure 5. Average wage estimates by NACE section and province

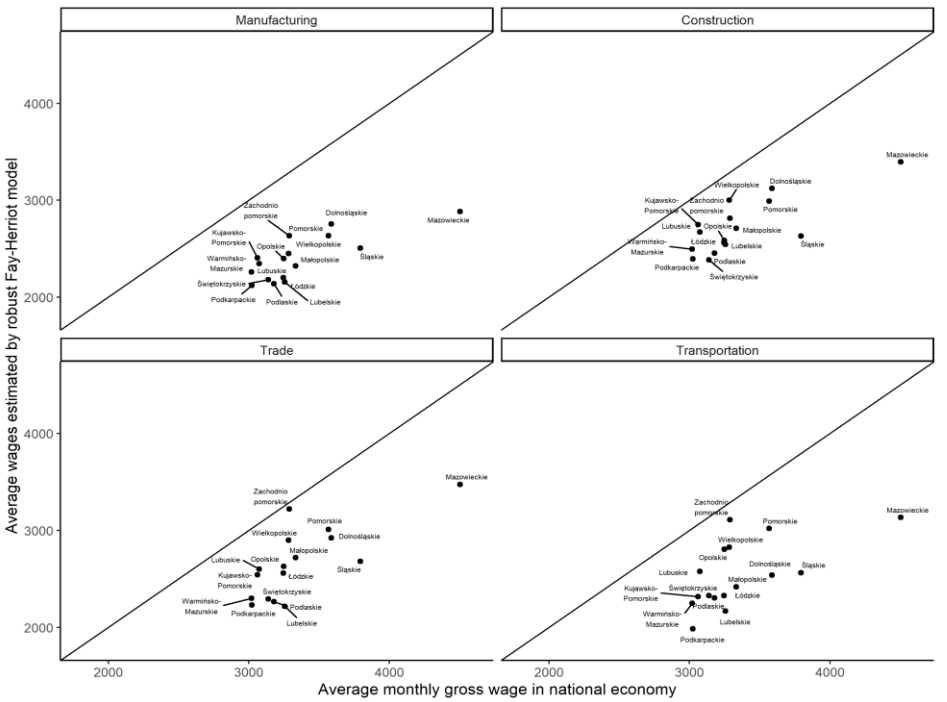


S o u r c e: based on data from the DG1 survey and the administrative register.

As Fig. 5 demonstrates, there is a strong spatial diversity in the average wage across provinces. The Mazowieckie province visibly stands out – average salaries reach the highest values in all the studied sections there. Average salaries reach the second highest values in the Dolnośląskie province (construction section) and in the Zachodniopomorskie (trade and transportation sections), whereas they assume the lowest values in Eastern Poland (in all the sections).

In the last part of the analysis, the obtained estimates are compared with the average gross wage in the national economy, which is presented in Fig. 6, in order to find out to what extent the estimates correspond with wage data from administrative registers.

Figure 6. Estimated average wage in small enterprices vs. average wage in the national economy by NACE sector in 2012



S o u r c e: based on data from the DG1 survey and the administrative register.

Fig. 6 shows a correlation between the estimates and the average wage in the national economy. Values of Pearson linear correlation coefficient vary from $r = 0.61$ for transportation to $r = 0.77$ for manufacturing. It is worth noting that the values of the average wage in the four sections are slightly lower than the national average.

5. ESTIMATION AT THE LOCAL LEVEL

As the estimation at the level of provinces (NUTS 2) was successfully conducted, the authors decided to carry out a similar estimation at the level, i.e. for districts (NUTS 3). Since there are many more territorial units at this level, the minimum sample size in particular domains was much smaller. In addition, there were some districts with no entities suitable for samples. As a result, calculations were made only for one section – manufacturing. Out of all the 379 districts, 350 were represented in the sample. The calculations yielded direct estimates of the average wage in small companies. These estimates ranged from 1258 PLN to 4246 PLN, while relative errors (RRMSE) ranged from 1% to 33%, with a mean of 11%.

After applying the robust Fay-Herriot model, the range of estimates did not change considerably – the minimum remained the same, while the maximum decreased to 3509 PLN. However, this method improved the estimation precision. The application of auxiliary variables made it possible to decrease the maximum RRMSE to 21%, with the mean at 8.7%. The above-described exercise shows that an average wage can also be estimated at the level of districts, but, given the smaller sample size in domains, this approach requires further analysis to test other sources of auxiliary information or other modified robust methods.

6. CONCLUSION

Indirect methods of estimation enable the estimation of the average wage for four NACE sections for the previously unpublished domains. The results obtained by means of the Fay-Herriot model and its robust version are, in most cases, more precise than the direct estimator when measured with the RRMSE. Moreover, robust estimation reduces the impact of outliers on the average wage and limits the range of estimates.

The results also show that the level of average wage varies across the four NACE sections. It assumes highest values in the Mazowieckie province. The size of bias was assessed using general data about the average monthly gross wage in the national economy.

It is worth noting that the application of the robust Fay-Herriot model at the level of districts has generally improved the estimation precision compared to direct estimation method. However, due to the fact samples are too small in some domains (even zero samples), there is a strong need for additional analysis to test other sources of auxiliary information, or other modified robust estimation methods.

REFERENCES

Boonstra H. J., Buelens B., (2011), *Model-based estimation*, Statistics Netherlands, Hague, Heerlen.

Dehnel G., (2017), GREG estimation with reciprocal transformation for a Polish business survey [in:] *Papież M., Śmiech S.* (eds.) *The 11th Professor Aleksander Zelias Internetional Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings*, Foundation of the Cracow University of Economics, Cracow, 67–75.

Dehnel G., Pietrzak M., Wawrowski Ł., (2017), Estymacja przychodu przedsiębiorstw na podstawie modelu Faya-Herriota, *Przegląd Statystyczny*, 64(1), 79–94.

Dehnel G., Wawrowski Ł., (2018), Robust estimation of revenues of Polish small companies by NACE section and province, [in:] Papież M., Śmiech S. (eds.), *Proceedings of the 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, Foundation of the Cracow University of Economics, Cracow, 110–119.

Eurostat (2013), *Handbook on precision requirements and variance estimation for ESS households surveys*, European Union, Luxembourg.

Fay R. E., Herriot R. A., (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74(366a), 269–277.

González-Manteiga W., Lombardia M. J., Molina I., Morales D., Santamaría L., (2008), Bootstrap mean squared error of a small-area EBLUP, *Journal of Statistical Computation and Simulation*, 78(5), 443–462.

GUS (2013), *Ludność. Stan i struktura demograficzno-społeczna. Narodowy Spis Powszechny Ludności i Mieszkań 2011*, GUS, Zakład Wydawnictw Statystycznych, Warszawa.

GUS (2017), *Działalność przedsiębiorstw niefinansowych w 2015 roku*, GUS, Warszawa.

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.

Huber P. J., (1981), *Robust Statistics*, John Wiley and Sons, New York.

Krzciuk, M. (2017). *On the Simulation Study of Jackknife and Bootstrap MSE Estimators of a Domain Mean Predictor for Fay-Herriot Model.* Acta Universitatis Lodziensis. Folia Oeconomica, *5*(331), 169-183.

Rao J. N. K., (2014), *Small-Area Estimation*, John Wiley & Sons, Hoboken, New Yersey.

Sinha S. K., Rao J. N. K., (2009), Robust small area estimation, *Canadian Journal of Statistics*, 37(3), 381–399.

Warnholz S., (2016), *Small Area Estimation Using Robust Extensions to Area Level Models,* (doctoral dissertation), Freie Universität, Berlin.

Żądło T., (2008), *Elementy statystyki małych obszarów z programem R*, Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego, Katowice.

Żądło T., (2012), O predykcji wartości globalnej w domenie z wykorzystaniem informacji o zmiennych dodatkowych przy założeniu modelu Faya-Herriota, *Acta Universitatis Lodziensis. Folia Oeconomica*, 271, 243–256.

DOI: 10.5604/01.3001.0013.9130

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Anna Gdakowicz[1]
Ewa Putek-Szeląg[2]
Wojciech Kuźmiński[3]

Examination of the effects of non-measurable explanatory variables on the value of real estate in the process of mass valuation of land[4]

Abstract. The paper proposes a solution to the problem of how to introduce non-measurable features (attributes) of a property that significantly affect its value to the process of its valuation. The authors adopt two measures enabling them to study the influence of order features on the value of property, the Spearman rank coefficients and standardized β_k coefficients, and proceed to check their efficiency, applying an algorithm of mass property valuation (SAMWN) to the sample of 567 plots of land in Szczecin designated for housing purposes. The results thus obtained are then compared with the valuations of these plots of land performed by property appraisers. The study demonstrates that lower valuation errors are obtained when using standardized β_k coefficients than the Spearman rank coefficients.

Keywords: method of statistical market analysis, mass valuation of real estate, non-measurable attributes, Spearman rank coefficient, standardized β_k coefficient

JEL: C35, R31

1. INTRODUCTION

According to the Real Estate Management Act[5], it is possible to estimate the market, the replacement and the cadastral values in the process of real estate valuation. The market value is defined as the most likely price that could be obtained for a given property at the date of valuation, under the following conditions: both parties to the transaction have to be independent of each other, have to be determined to enter into the deal, have to act of their free will and have to

[1] University of Szczecin, Institute of Economics and Finance, ul. Mickiewicza 64, 71–101 Szczecin, ORCID: https://orcid.org/0000-0002-4360-3755, corresponding author – e-mail: anna.gdakowicz@usz.edu.pl.

[2] University of Szczecin, Institute of Economics and Finance, ul. Mickiewicza 64, 71–101 Szczecin, ORCID: https://orcid.org/0000-0002-9302-2115.

[3] University of Szczecin, Institute of Economics and Finance, ul. Mickiewicza 64, 71–101 Szczecin, ORCID: https://orcid.org/0000-0003-3256-9093.

[4] Article financed under the project of the National Science Centre, registration no. 2017/25/B/HS4/01813.

[5] Act of 21 August 1997 on Real Estate Management, Journal of Laws of 1997, No 115, Item 741, as amended.

have the same knowledge about the property. Additionally, the property has to be exposed to the market for a sufficient period of time. The market value can only be determined for properties that are tradable. Replacement value is determined as the estimated amount consisting of the cost of the acquisition of land (its market value) and the cost of the production of property components, taking into account the degree of wear and tear and assuming that production costs were incurred at the date of valuation. Unlike the market value, the replacement value applies to properties which are not tradable (due to the type of property, its current use or purpose). The cadastral value is determined during the general taxation, however, there are no legal acts which would specify in a detailed way the manner and methodology for determining this value.

The person authorised to determine the value of real estate in Poland is a property appraiser. Such person is responsible for choosing an optimal way of estimating the value of a particular property. This value may be estimated using the comparative, income-based, mixed or cost-based approach (Tabl. 1), depending on the type of real estate, the type of value to be determined, the purpose of the valuation and the availability of data.

TABLE 1. LIST OF APPROACHES, METHODS AND TECHNIQUES OF PROPERTY VALUATION AND THE TYPE OF THE PROPERTY VALUE ASSESSED

Approach	Method	Technique	Type of property value to be assessed
Comparison	pair sales comparison	–	market value
	mean price correction		
	market statistical analysis		
Income	investment	simple capitalisation	
		discounting flows income	
	profit	simple capitalisation	
		discounting flows income	
Mixed	residual	–	
	liquidation costs	detailed	
		index	
	land estimate indexes	–	
Cost	replacement costs	detailed	replacement value
		joined elements	
		index	
	substitution costs	detailed	
		joined elements	
		index	

S o u r c e: own compilation on the basis of the Regulation of the Council of Ministers of 21 September 2004 on the valuation of real estate and preparation of a valuation report (Journal of Laws of 2004, No 207, Item 2109).

Almost all valuation methods and techniques are described and explained in detail in the Real Estate Valuation Regulation[6] and in the national valuation standards of the Polish Federation of Valuers' Associations[7]. There is one exception, however, namely the method of the statistical analysis of the market, which has not been described in detail by any legal act. The current legislation refers[8] to the idea of a comparative approach, which involves comparing the property under valuation with similar properties traded on the market. The method of the statistical analysis of the market necessitates not only using a suitable set of transaction prices of similar properties as a reference, but also, in order to achieve more precise results, obtaining information about the terms of these transactions and on the characteristics of these properties that affect their prices. It is therefore necessary to build a database of properties similar to the property under valuation and to define a set of attributes for each of them. The attributes may vary depending on the type of property and on the type of the market. The Polish legislator and valuers' associations have left the choice of further calculation procedures to the person carrying out the valuation. There are no guidelines as to the appropriate algorithm, pattern of conduct or conditions enabling the adoption of suitable statistical-econometric methods

At the turn of the 20th and 21st century, many attempts were made to apply an econometric model (most often in the form of linear regression models) to estimations of the price or value of real estate (Czaja and Żak, 1993; Sawiłow, 1995; Cellmer, 1999; Źróbek and Bełej, 2000; Dacko, 2000a; Dacko, 2000b; Cellmer, 2000; Lipieta, 2000; Pawlukowicz, 2001; Zadumińska and Sztaudynger, 2001; Hopfer et al., 2001; Źróbek, 2002; Lis, 2008; Renigier-Biłozor, 2008; Zbyrowski, 2010; Walkowiak and Zydroń, 2012; Doszyń and Gnat, 2016; Kubus, 2016; Gdakowicz and Putek-Szeląg, 2018)[9]. In those models, properties were described by various features, for example: their purpose determined by local zoning plans, available utilities, accessibility, size, shape of the plot, type of buildings, year of construction and quality of the land or landscape. The number of the variables was limited only by the availability of data. With the expansion of databases and the general improvement of computer skills among property valuers and real estate analysts, the matching parameters of estimated econometric models have improved. Those models were not subject to a reliable statistical verification; most likely, it was only the validity of their structural parameters and the degree to which they fitted which were checked.

[6] The Regulation of the Council of Ministers of 21 September 2004 on the valuation of real estate and preparation of a valuation report (Journal of Laws of 2004, No 207, Item 2109).

[7] *National Valuation Standards of the Polish Federation of Valuers' Associations*, https://pfsrm.pl/aktualnosci/item/14-standardy-do-pobrania, (accessed 20.08.2018).

[8] The Regulation of the Council of Ministers of 21 September 2004, op. cit.

[9] In the article, the authors refer to articles written by Polish authors, because the proposed models take into consideration the Polish legal status. In other countries, the problem of mass valuation using econometric and statistical methods was also discussed, e.g. Kauko and D'Amato (2008); Yasnitsky and Yasnitsky (2016); Arribas et al. (2016); Ciuna et al. (2017).

The method of statistical market analysis is likely to gain additional significance in the context of the increasing demand for mass valuation of real estates. Mass valuation applies when (Hozer et al., 1999; Kuryj, 2007; Telega et al., 2002):

- the subject of the valuation is a large number of properties of one type,
- the valuation is carried out by means of a uniform, objective approach which yields consistent results,
- all properties subject to the valuation are assessed simultaneously, i.e. on the basis of data (the state of the property and the level of prices) collected on the same day for all the valued properties.

Due to the scale of the process, classical and non-classical multidimensional methods of statistical analysis tend to be applied while searching for practical solutions – for example, in general taxation (Benjamin et al., 2004; Kauko and D'Amato, 2008). However, when using these methods, it is not possible to fully utilize non-measurable explanatory variables (which influence the value of properties – the so-called real estate attributes) in the analysis. So far, the impact of individual attributes (often encoded as features measured on ordinal scales) on the value of estate property has been measured using the Pearson correlation coefficient (Czaja and Dąbrowski, 2008; Czaja and Ligas, 2010), Spearman rank correlation coefficient (Gaca and Sawiłow 2014; Babatunde, 2018; Gaca, 2018) or conjoint analysis (Pawlukowicz and Bartłomowicz, 2005; Głuszak, 2011).

The question arises whether it is possible to measure the influence of variables on the value of a property when it is not possible to measure their quantitative condition, as for example in the case of the current market trend concerning location. An immediate answer to this question would be negative, because if we do not observe a variable, we cannot measure its effect, and if we do not measure the effect, we cannot measure its impact. However, the empirical study presented in this paper demonstrates the opposite – indeed, it is possible to measure the effects of such variables. The aim of the paper, therefore, is twofold: to assess the influence of non-measurable features (attributes) on the value of real estate with the use of Spearman rank coefficients and the standardized β_k coefficients, as well as utilizing these features in the process of real estate valuation, conducted according to the Szczecin algorithm of mass real estate valuation (SAMWN).

2. HOW TO MEASURE THE NON-MEASURABLE?

The analysis of the real estate market demonstrates that the location of a property is one of the attributes which strongly influence its value. A residential property located in an attractive, fashionable district will be valued

higher than a similar property located in an unattractive area, far from the city centre. Location is a qualitative feature. Experts try to quantify this attribute by describing it as desirable, average or undesirable. But even this kind of definition is very subjective – the assessment of the attractiveness of a location depends, at least to some extent, on the emotions and the potential associations the person describing the property might have with a given location. So, how to measure the effect of this qualitative variable on the value of a property? Guzik (2008) proposed an approach where the attractiveness coefficient for particular locations, i.e. the location rent, is incorporated into the econometric model.

In the econometric analysis, when examining the relationship:

$$X_{1t} = f(X_{2t}, X_{3t}, \dots, X_{kt}, U_t) \tag{1}$$

we can use, e.g.:

1. levels of variables X_{it},
2. changes $\Delta X_{it} = X_{it} - X_{it-1}$,
3. effects of variables $X_{2t}, X_{3t}, \dots, X_{kt}$ on X_{1t} (structural parameters),
4. outcome of effects of variables X_{it}, i.e. $X_{1t}(X_{it}); i = 2, 3, \dots, k$.

It appears that even when it is not possible to examine the levels and relations listed in points 1 to 3, we can still examine the effects of non-measurable explanatory variables (attributes) on the explanatory variable (Hozer, 2003). In order to be able to examine the effects referred to in point 4, it is necessary to conduct a special procedure based on the non-classical model of relationships described in point 2.

3. METHODOLOGY

In the first phase of the study, variables that significantly affect the value of a property were specified, out of which these attributes were selected that both had the strongest effect on the value of a property and at the same time were collectable, e.g.: size, transport accessibility, neighbourhood, development, utilities, land and water conditions. It is often impossible to meet both of these conditions simultaneously. The Szczecin land property mass valuation algorithm (SAMWN) presented below, however, takes into account both the deliberate human activity and non-measurable factors in the form of the market value coefficients (WWR_j):

$$\widehat{W}_{ji} = WWR_j \cdot pow_i \cdot W_{baz} \cdot \prod_{k=1}^{K}(1 + A_k), \tag{2}$$

where:

$\widehat{W}_{ji}$ – market (cadastral) value of the i-th property in the j-th elementary area,
WWR_j – market value coefficient in the j-th elementary area $(j = 1, 2, \dots, J)$,
J – number of elementary areas,
pow_i – size of the i-th property,
W_{baz} – price of 1 m^2 of the cheapest land in the valuated area,
A_k – effect of the k-th attribute $(k = 1, 2, \dots, K)$,
K – number of attributes.

Coefficients WWR_j are computed for individual elementary areas[10] as an arithmetical mean of the WWR_i (formula 3) calculated for individual properties-representatives of each of the elementary areas. These, in turn, are the quotients of the market value of the property (formula 4) determined by the property valuer[11] in the process of individual valuation and the hypothetical value of the property determined on the basis of formula 5:

$$WWR_j = \frac{\sum_{i=1}^{l} WWR_i}{l}, \tag{3}$$

$$WWR_i = \frac{WR_{ri}}{\widehat{W}_{hi}}, \tag{4}$$

$$\widehat{W}_{hi} = pow_i \cdot W_{baz} \cdot \prod_{k=1}^{K}(1 + A_k), \tag{5}$$

where:

WWR_i – ratio of the market value to the hypothetical value of the i-th property,
l – number of properties-representatives in the j-th elementary area,
WR_{ri} – market value of the i-th property, determined by a property valuer,
$\widehat{W}_{hi}$ – hypothetical value of the property calculated on the basis of the model.

In the SAMWN algorithm (formula 2), it is problematic to determine the A_k coefficients whose function is to measure to what extent particular attributes (features) affect the value of a property. Since the attributes are presented on an ordinal scale, the following two methods have been used to determine the effects of particular characteristics on the value of properties: the Spearman coefficients (R_{xy}) and standardised β_k coefficients. The latter are calculated according to the following formula:

[10] The elementary area is defined as an area in which a certain number of valued properties are located that are characterised by the same effect of the location attribute on their value.

[11] Property valuers who estimated the value of given properties used location as one of attributes describing the property.

$$\hat{\beta}_k = \frac{S_{A_k}}{S_{WR_r}} \cdot \frac{(WR_{ri} - \overline{WR}_r)}{(A_k - \bar{A}_k)}, \tag{6}$$

where:

$\hat{\beta}_k$ – standardised beta coefficient of the k-th attribute,
S_{WR_r} – standard deviation of the value of 1 m^2 of land determined by a property valuer
$\overline{WR}_r$ – average value of 1 m^2 of land calculated on the basis of values determined by a property valuer,
S_{A_k} – standard deviation of the effect of the k-th attribute,
$\bar{A}_k$ – average value of the effect of the k-th attribute

The calibration of the attributes of land properties is carried out on the basis of correction coefficients $(1 + A_k)$, which are determined according to the method of distance from extreme values (Lis, 2008):

$$\begin{aligned} 1 + A_k &= \left(1 - \frac{1}{2}\rho\right) + \left[\left(1 + \frac{1}{2}\rho\right) - \left(1 - \frac{1}{2}\rho\right)\right] \cdot \frac{l_{kp}}{k_p - 1} \\ &= \left(1 - \frac{1}{2}\rho\right) + \rho \frac{l_{kp}}{k_p - 1}, \end{aligned} \tag{7}$$

where:

l_{kp} – the p-th category of the k-th attribute,
k_{p} – number of categories of the k-th attribute,
ρ – standardised coefficients of the k-th attribute, depending on the method adopted: Spearman coefficient R_{xy} or beta coefficient $\hat{\beta}_k$.

In order to fully explain the value of the property (in 100%), the estimates of the relevant Spearman coefficients and standardised beta coefficients are adjusted, so that the sum of their absolute values equals one.

At the next stage of the study, the results of property valuations carried out by individual valuers are juxtaposed with the results obtained through SAMWN, using the adjusted Spearman and beta coefficients.

The results thus obtained are compared using a relative valuation error:

$$\partial = \sum_{i=1}^{n} \frac{\left|W_{ji} - WR_{ri}\right|}{W_{ji}} \cdot 100\% \tag{8}$$

and the following two variation measures:

$$Se = \sqrt{\frac{\sum_{i=1}^{n}\left(WR_{ri} - WR_{ji}\right)^2}{n}}, \tag{9}$$

$$Vs = \frac{Se}{\overline{WR_{ri}}} \cdot 100\%, \tag{10}$$

where:
Se – standard deviation of the value of 1 m^2 land,
Vs – variation coefficient of the value of 1 m^2 of land.

4. EMPIRICAL EXAMPLE

The study used the data on 567 plots of land in Szczecin designated for housing purposes, which were the subject of individual valuation in 2005. The plots were located in 5 elementary areas (Tabl. 2).

TABLE 2. NUMBER OF INDIVIDUAL ELEMENTARY AREAS COVERED BY THE STUDY

Elementary area	Number
3	187
4	37
5	178
6	62
7	103
Total	567

S o u r c e: own compilation.

The plots were described by the following attributes:
y – value of 1 m^2 (in PLN), a dependent variable,
x_1 – physical traits (0 – undesirable, 1 – average, 2 – desirable),
x_2 – development (0 – no, 1 – yes),
x_3 – utilities (0 – no, 1 – partial, 2 – full),
x_4 – neighbourhood (0 – undesirable, 1 – desirable),
x_5 – accessibility (0 – poor, 1 – average, 2 – good),
x_6 – location (0 – undesirable, 1 – average, 2 – desirable),
x_7 – size (0 – large, 1 – medium, 2 – small),
x_8 – ground and water conditions (0 – bad, 1 – undesirable, 2 – average, 3 – desirable).

Since the main purpose of the paper is to present the method of calculating the effect of non-measurable variables on the value of real estate, the location attribute was omitted in the subsequent calculations. The value of this attribute was determined on the basis of the opinion of a property valuer, who while deciding about it, took into account the popularity of the given area. The estimates of the Spearman correlation coefficients and $\hat{\beta}_k$ coefficients between the value of 1 m^2 of a land property in Szczecin and individual attributes are presented in Tabl. 3

TABLE 3. ESTIMATES OF THE SPEARMAN CORRELATION COEFFICIENT AND $\hat{\beta}_k$ COEFFICIENTS OF 1 M² AND INDIVIDUAL ATTRIBUTES OF LAND PROPERTIES IN SZCZECIN IN 2005

Coefficient	x_1	x_2	x_3	x_4	x_5	x_7	x_8
R_{xy}	–0.063	**0.282**	**0.343**	–0.074	**0.175**	–0.081	**0.187**
Adjusted R_{yx}		**0.286**	**0.347**		**0.177**		**0.190**
$\hat{\beta}_k$	0.039	**0.106**	**0.158**	–0.049	**0.092**	**–0.155**	**0.389**
Adjusted $\hat{\beta}_k$		**0.118**	**0.176**		**0.102**	**–0.172**	**0.433**

x_1 – physical traits, x_2 – development, x_3 – utilities, x_4 – neighbourhood, x_5 – accessibility, x_7 – size, x_8 – water and ground conditions.
Figures in bold – significant at 5%.
S o u r c e: own calculation.

They indicate that when determining the impact of attributes using the adjusted Spearman coefficients, the physical traits, neighbourhood and size variables turned out to be insignificant; whereas when using the standardised beta coefficients, only physical traits and neighbourhood were insignificant.

According to the Spearman coefficients, the value of properties was influenced by utilities to the largest extent. In the case of beta coefficients, the highest correlation was observed for the land and water conditions. All the coefficients had low magnitudes.

Adjusted coefficients were calculated by adjusting the significant values of the coefficients of individual attributes, so that their sum equalled one. Only the attributes significantly affecting the value of the property were taken into account.

Tabl. 4 shows the calculation of the effect of each attribute on the value of the property.

TABLE 4. CALCULATION OF VALUES OF LAND PROPERTY ATTRIBUTES

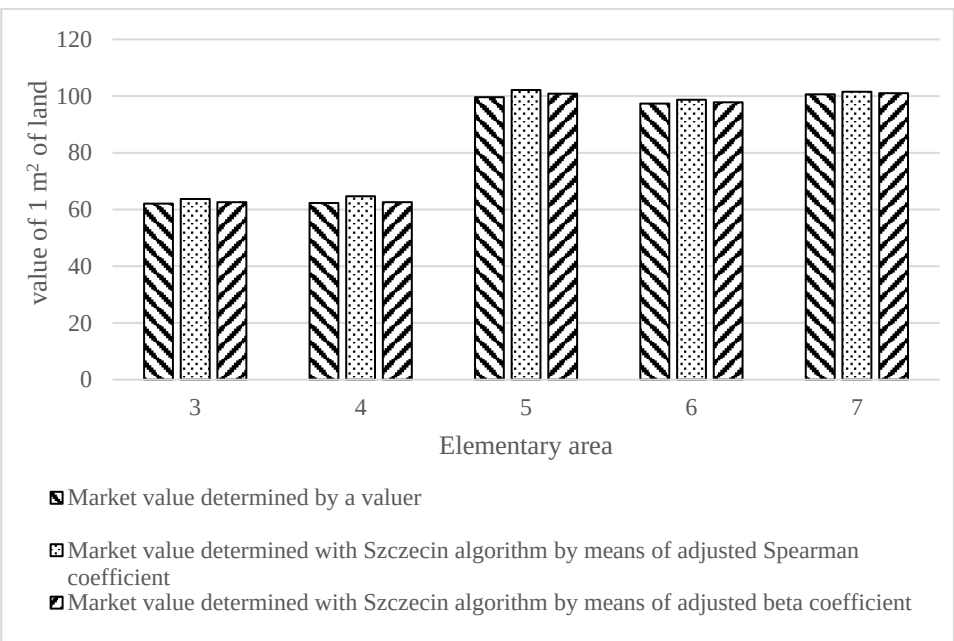
Attribute	Attribute alternative	Adjusted R_{xy}	$1 + A_k$	A_k %	Adjusted $\hat{\beta}_k$	$1 + A_k$	A_k %
Development	0	0.286	0.857	–14.29	0.118	0.941	–5.9
	1		1.143	14.29		1.000	0
Utilities	0	0.347	0.827	–17.35	0.176	0.912	–8.79
	1		1.000	0		1.000	0
	2		1.174	17.35		1.088	8.79
Accessibility	0	0.177	0.911	–8.86	0.102	0.949	–5.08
	1		1.000	0		1.000	0
	2		1.089	8.86		1.051	5.08
Size	0	–	–	–	–0.172	1.086	8.6
	1		–	–		1.000	0
	2		–	–		0.914	–8.6
Ground and water conditions	0	0.190	0.905	–9.49	0.433	0.784	–21.63
	1		0.968	–3.16		0.928	–7.21
	2		1.032	3.16		1.072	7.21
	3		1.095	9.49		1.216	21.63

S o u r c e: own calculations.

The power of the effect of the attributes on the value of a property varies depending on the applied coefficient. When we use the adjusted Spearman coefficient, it is utilities that affect the value of 1 m^2 of land to the largest extent. Plots equipped with all the required utilities are on average 34.7% more expensive than non-equipped plots. The second most influential feature is development. The attributes which relatively have the smallest effect on the value of the property are ground and water conditions and accessibility.

On the other hand, when applying the adjusted $\hat{\beta}_k$coefficient, ground and water conditions turned out to be a variable shaping the value of a property to the largest extent. A plot of land with favourable ground and water conditions was on average 43.3% more expensive than a plot with poor such conditions. All attributes affected the value of the property, including size, however, what is questionable here, is the sign of the correlation – the smaller the plot, the lower the value of 1 m^2 (1 m^2 of a small plot was worth 17.2% less than 1 m^2 of a large plot). Interestingly, the observation of the real estate market shows something opposite, namely positive rather than negative correlation, i.e. the smaller the plot, the higher the value (price) of 1 m^2 (Foryś and Gdakowicz, 2004). This inconsistency might result from the fact that small plots of land belonged to natural persons (and the value of the plots was lower), while large plots were owned by institutionalised entities, and the value of these properties was higher.

Figure 1. Summary of the average value of 1 m^2 of land estimated by property valuers and calculated using SAMWN with the application of adjusted Spearman coefficients and standardised beta coefficients in individual elementary areas



S o u r c e: own calculations.

The estimations of the average value of 1 m^2 of land carried out both by property valuers and using the Szczecin mass valuation algorithm (with the use of both approaches) yielded similar results, in each of the elementary areas (Fig-ure 1). According to property valuers, popular and attractive plots (i.e. worth more) were located in elementary areas No 5, 6 and 7, where the value of 1 m^2 of the plot reached about 100 PLN. The application of the Szczecin algorithm of mass valuation of real estate confirmed the above results – plots located in areas 5, 6 and 7 were valued higher than plots located in areas 3 and 4. The application of the SAMWN calculation algorithm and the estimation of WWR_j values for particular elementary areas enabled the inclusion of the effect of the plot location (fashion) in the calculation, although that variable was not one of the *a priori* attributes.

Tabl. 5 presents values of market coefficients (WWR_j) estimated for particular elementary areas using SAMWN. The results obtained through the application of the algorithm (in two variants: using the adjusted Spearman and beta coefficients) are compared with the values estimated by property valuers. The consecutive columns present measures of agreement between the obtained results, such as the residual deviation, coefficient of variation and relative valuation error.

TABLE 5. COEFFICIENTS OF MARKET VALUES FOR PARTICULAR ELEMENTARY AREAS AND MEASURES OF AGREEMENT BETWEEN SAMWN RESULTS AND VALUERS' APPRAISALS

Elementary area	Adjusted R_{xy}				Adjusted $\hat{\beta}_k$			
	WWR_j	Se	Vs	∂	WWR_j	Se	Vs	∂
3	0.978	8.377	13.50	13.03	0.983	4.643	7.48	6.05
4	0.987	10.525	16.89	16.96	0.973	3.507	5.63	4.54
5	1.546	13.736	13.79	13.13	1.575	9.226	9.26	6.73
6	1.537	9.641	9.91	7.73	1.431	5.978	6.15	4.33
7	1.449	7.663	7.61	6.12	1.546	5.432	5.40	4.45

S o u r c e: own calculations.

The coefficient of the market value in No 5 elementary area (for the Spearman coefficients) is 1.546, which means that the value of land in this area, as calculated with the use of SAMWN, was on average 54.6% higher than the value of land located in a less attractive elementary area. The same coefficient in the same elementary area for the $\hat{\beta}_k$ coefficients, however, totals 1.575, which means that the value of land in this area, also calculated via SAMWN, is on average 57.5% higher than the value of land located in a less fashionable area.

When the SAMWN with the adjusted Spearman coefficient was applied, the value of a plot of land in No 3 elementary area differed from the value estimated by the property valuer on average by +/– 8.38 PLN per 1 m^2, which constituted 13.5% of the average value of land determined by the valuer. When having applied the adjusted $\hat{\beta}_k$ coefficient for the same elementary area, however, the value of 1 m^2 of land differed on average by +/– 4.64 PLN per 1 m^2 from the val-

ue estimated by the valuer, which equals 7.48% of the average value of land determined by the valuer.

The analysis of the results of the relative valuation error shows a smaller discrepancy between the results of the valuation carried out by valuers and the results obtained with SAMWNs using the adjusted beta coefficients than when using the adjusted Spearman coefficients. In all the elementary areas the results thus obtained showed lower values of stochastic structure parameters.

5. CONCLUSION

Taking account of the impact of non-measurable variables on the dependent variable proves particularly useful for real estate market analysts. Many attributes that influence the value and price of a property are non-measurable, for example fashion, attractiveness or popularity. The article proposes a procedure for estimating the value of a property in the form of mass valuation, in which the attributes related to location and fashion are not included *a priori*.

The values of properties estimated using the SAMWN and those obtained on the basis of individual valuers' appraisals turned out to be similar. The construction of the algorithm makes it possible, through the estimation of the WWR_j, to take the degree of impact non-measurable attributes have on the value of the property into account while performing the calculation. Among the two proposed methods of determining the influence of attributes on the value of a property, better results were obtained when adjusted beta coefficients were applied.

The proposed method for estimating the value of a property has assumed particular importance in the context of increasing demand for mass valuation of real estate and the method of statistical market analysis. The legislator has not defined a detailed procedure for any of these approaches, thus leaving many decisions to the discretion of property valuers. This study may therefore be an important voice in the debate on the use of econometric and statistical methods in the process of real estate valuation.

REFERENCES

Arribas I., Garcia F., Guijarro F., Oliver J., Tamosiuniene R., (2016), Mass Appraisal of Residential Real Estate Using Multilevel Modelling, *International Journal of Strategic Property Management*, 20(1), 77–87, DOI: 10.3846/1648715X.205.1134702.

Babatunde I. O., (2018), Examining Heuristics for Building – Work-In-Progress Valuations in Niger State Nigeria, *Real Estate Management and Valuation*, 26(2), 92–103, DOI: 10.2478/remav-2018-0019.

Benjamin J. D., Randall S., Guttery R. S., Sirmans C. F., (2004), Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation*, Journal of Real Estate Practice and Education*, 7(1), 65–77.

Cellmer R., (1999), *Zasady i metody analizy elementów składowych rynku nieruchomości*, Wyd. Educaterra, Olsztyn.

Cellmer R., (2000), Metody analizy rynku nieruchomości – przykłady, *Wycena*, 2(49), 6–10.

Ciuna M., Milazzo L., Salvo F., (2017), A Mass Appraisal Model Based on Market Segment Parameters, *Buildings*, 7, 34, DOI: 10.3390/buildings7020034.

Czaja J., Dąbrowski J., (2008), Special Algorithms for Assesing Market Value of Real Estates, *Geomatics and Environmental Enginering*, 2(2), 21–31.

Czaja J, Ligas M., (2010), Zaawansowane metody analizy statystycznej rynku, *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 18(1), 7–19.

Czaja J., Żak M., (1993), *Systemy przetwarzania danych przy szacowaniu nieruchomości metodami rynkowymi*, Acta Acad. Agricult. Tech. Olst., 451, Geod. Ruris Regulat., 24: 7–19.

Dacko M., (2000a), Solver – zastosowanie w modelowaniu ekonometrycznym na potrzeby analiz rynku nieruchomości, *Wycena*, 4, 41–45.

Dacko M., (2000b), Zastosowanie regresji wielokrotnej w szacowaniu nieruchomości w arkuszu kalkulacyjnym Microsoft Excell 2000, *Wycena*, 2, 50–56,

Doszyń M., Gnat S., (2016), Taksonomiczno-ekonometryczna procedura wyceny nieruchomości dla różnych miar porządkowania, *Prace Naukowe Uniwersytetu ekonomicznego we Wrocławiu nr 427. Taksonomia 27 Klasyfikacja i analiza danych – teoria i zastosowania*, 84–93.

Foryś I., Gdakowicz, (2004), Wykorzystanie metod ilościowych do badania rynku nieruchomości, *Studia i materiały Towarzystwa Naukowego Nieruchomości*, 12(1), 41–49.

Gaca R., Sawiłow E., (2014), Zastosowanie współczynnika korelacji rang Spearmana do ustalenia wag cech rynkowych nieruchomości, *Rzeczoznawca majątkowy*, 82, 24–30.

Gaca R., (2018), Parametric and Non-Parametric Statistical Methods in the Assessment of the Effect of Property Attributes on Prices, *Real Estate Management and Valuation*, 26(2), 83–91, DOI: 10.2478/remav-2018-0018.

Gdakowicz A., Putek-Szeląg E., (2018), Ekonometryczno-statystyczne metody masowej wyceny nieruchomości w Polsce – studium przypadków, W: Nieruchomość w Przestrzeni 4, Wydawnictwa Uczelniane Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy, Bydgoszcz, 141–154.

Głuszak M., (2011), Eksperyment conjoint jako metoda wyznaczania współczynników wagowych atrybutów w wycenie nieruchomości, *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, (850), 49–64.

Guzik B., (2008), Statystyczne metody szacowania atrakcyjności lokalizacji mieszkań Statistical Methods for Estimating the Attractiveness of Apartments Location, *Roczniki Ekonomiczne Kujawsko-Pomorskiej Szkoły Wyższej w Bydgoszczy*, (1), 233–250.

Hopfer A., Jędrzejewski H., Źróbek R., Źróbek S., (2001), *Podstawy wyceny nieruchomości*. Twiger, Warszawa.

Hozer J., (2003), *Tempus locus homo casus et fortuna regit factum. Zbiór esejów ekonomicznych*, Oficyna ”in Plus”, Szczecin.

Hozer J., Foryś I., Zwolankowska M., Kokot S., Kuźmiński W., (1999), *Ekonometryczny algorytm masowej wyceny nieruchomości gruntowych*, Stowarzyszenie Pomoc i Rozwój, Szczecin.

Kauko T., D'Amato M., ed., (2008), *Mass Appraisal Methods: An International Perspective for Property Valuers,* Blackwell Publishing Ltd.

Kubus M., (2016), Locally Regularized Linear Regression in the Valuation of Real Estate, *Statistics in Transition new series*, 17(3), 515–524

Kuryj J., (2007), Metodyka wyceny masowej nieruchomości na bazie aktualnych przepisów prawnych, *Wycena*, 4(81), 50-58.

Lipieta A., (2000), Model ekonometryczny ze zmiennymi jakościowymi opisujący ceny mieszkań, *Wiadomości Statystyczne*, 8, 10–20.

Lis Ch., (2008), Wykorzystanie metod ilościowych w procesie powszechnej taksacji nieruchomości w Polsce, *Prace Naukowe Akademii Ekonomicznej w Katowicach, Modelowanie matematyczne i ekonometryczne na polskim rynku finansowym*, 191–204.

National Valuation Standards of the Polish Federation of Valuers' Associations, Standardy – do pobrania, https://pfsrm.pl/aktualnosci/item/14-standardy-do-pobrania, (accessed: 20.08.2018).

Pawlukowicz R., (2001), Przegląd propozycji określania wartości rynkowej nieruchomości z wykorzystaniem modeli ekonometrycznych, *Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 320, Prace naukowe Katedry Ekonometrii i Statystyki Szczecin*, 317–324.

Pawlukowicz R., Bartłomowicz T., (2005), Conjoint analysis jako sposób wyznaczania wag cech rynkowych w wycenie rynkowej nieruchomości za pomocą podejścia porównawczego, *Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 1096, Ekonometria, 15*, 128–139.

Renigier-Biłozor M., (2008), Zastosowanie teorii zbiorów przybliżonych do masowej wyceny nieruchomości na małych rynkach, *Acta Scientiarum Polonorum, Administratio Locorum*, 7(3), 35–51.

Sawiłow E., (1995), *Próba matematycznego modelowania wartości gruntów na terenach zurbanizowanych*, Wyd. Stowarzyszenie Rzeczoznawców Majątkowych we Wrocławiu i Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, Wrocław, Warszawa.

Telega T., Bojar Z., Adamczewski Z., (2002), Wytyczne przeprowadzeni powszechnej taksacji nieruchomości, *Przegląd Geodezyjny*, 6, 6–11.

Walkowiak R., Zydroń A., (2012), Zastosowanie regresji krokowej do określenia atrybutów wpływających na wartość nieruchomości rolnych na przykładzie gminy Mosina, *Acta Scientiarum Polonorum Administratio Locorum*, 11(3), 239–253.

Yasnitsky L. N., Yasnitsky V. L., (2016). Technique of Design for Intergrated Economic and Mathematical Model for Mass Appraisal of Real Estate Property. Study Case of Yekaterinburg Housing Market, *Journal of Applied Economic Sciences*, 11(46), 1519–1531.

Zadumińska M., Sztaudynger J., (2001), Wykorzystanie modelu ekonometrycznego do wyceny nieruchomości, *Wiadomości Statystyczne*, 1, 6–13.

Zbyrowski R., (2010), Szacowanie wartości nieruchomości na podstawie modeli ekonometrycznych, *Equilibrium*, 1(4), 241–252.

Źróbek S., (2002), *Określenie wartości rynkowej nieruchomości*, Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego, Olsztyn.

Źróbek S., Bełej M., (2000), *Podejście porównawcze w szacowaniu nieruchomości*, Wyd. Educaterra, Olsztyn.

DOI: 10.5604/01.3001.0013.9131

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Mariusz ŁAPCZYŃSKI[1]
Bartłomiej JEFMAŃSKI[2]

The number of clusters in hybrid predictive models: does it really matter?

Abstract. *For quite a long time, research studies have attempted to combine various analytical tools to build predictive models. It is possible to combine tools of the same type (ensemble models, committees) or tools of different types (hybrid models). Hybrid models are used in such areas as customer relationship management (CRM), web usage mining, medical sciences, petroleum geology and anomaly detection in computer networks. Our hybrid model was created as a sequential combination of a cluster analysis and decision trees. In the first step of the procedure, objects were grouped into clusters using the k-means algorithm. The second step involved building a decision tree model with a new independent variable that indicated which cluster the objects belonged to. The analysis was based on 14 data sets collected from publicly accessible repositories. The performance of the models was assessed with the use of measures derived from the confusion matrix, including the accuracy, precision, recall, F-measure, and the lift in the first and second decile. We tried to find a relationship between the number of clusters and the quality of hybrid predictive models. According to our knowledge, similar studies have not been conducted yet. Our research demonstrates that in some cases building hybrid models can improve the performance of predictive models. It turned out that the models with the highest performance measures require building a relatively large number of clusters (from 9 to 15).*

Keywords: hybrid predictive model, k-means algorithm, decision trees

JEL Classification: C10, C18, C52

1. INTRODUCTION

The aim of this paper is to check to what extent the number of clusters affects the quality of predictive models which combine decision trees with cluster analysis (centre-based algorithm). The concept of the hybridization of the two methods is not new – it was already applied to customer relationship management (Chu et al., 2007; Bose and Chen, 2009; Li et al., 2011), the analysis of the Internet users' patterns of behaviour (Łapczyński and Surma, 2012), medical sci-

[1] Cracow University of Economics, College of Management and Quality Sciences, Department of Market Analysis and Marketing Research, ul. Rakowicka 27, 31-510 Cracow, Poland, corresponding author – e-mail: lapczynm@uek.krakow.pl, ORCID: https://orcid.org/0000-0002-4508-7264.

[2] Wroclaw University of Economics and Business, Faculty of Economics and Finance, Department of Econometrics and Informatics, ul. Nowowiejska 3, 58-500 Jelenia Góra, Poland, ORCID: https://orcid.org/0000-0002-0335-0036.

ences (Khan and Mohamudally, 2011; Shouman et al., 2012), petroleum geology (Ferraretti et al., 2011) and the detection of computer anomalies (Gaddam et al., 2007). Hybrid models differ from ensemble models in that they combine two different analytical tools. In the case of ensemble models, the commonly used procedures include the random forest, rotation forest, and boosted trees. Hybrid models, on the other hand, are referred to as cascade models, cross-algorithm ensembles, and two-stage classification (Łapczyński and Jefmański, 2013).

We decided to combine a popular decision tree algorithm CART (classification and regression trees) with the *k*-means algorithm. The hybridization process was tested on 14 data sets downloaded from publicly accessible online repositories. Each of them had a qualitative dependent variable with two or more categories, and a set of independent variables presented on various measurement scales. In addition, we calculated 4 cluster validity measures. However, our primary objective was to analyse hybrid models based on 2 to 20 clusters.

The paper consists of 4 sections. Section 2 encompasses a brief description of the employed analytical tools and the method for building a hybrid model. It also presents the characteristics of data sets and the description of the process of data preparation. Section 3 discusses the results of the study and provides the assessment of the quality of hybrid models based on five performance measures. The authors also explain there why hybrid models demonstrate a higher predictive power for some of the data sets than for the others. The conclusions and recommendations are provided in the last section.

2. THE CHARACTERISTICS OF A HYBRID MODEL AND EMPLOYED DATA SETS

2.1 CART – *k*-MEANS HYBRID MODEL

A decision tree is a commonly used analytical tool for data mining. The analysis utilises the CART algorithm, developed by Breiman et al. (1984). This tool demonstrates great flexibility in terms of the measurement scale of independent variables. It does not have such a great predictive power as ensemble models, but it enables creating a set of rules according to an 'if ... then ...' formula, which is easy to understand for managers with no mathematical background. The analysis adopts the CART algorithm where equal a priori probabilities and equal misclassification costs have been assumed. A minimum number of cases in tree leaves is placed at the level of 2% of the training set.

A cluster analysis with the use of the *k*-means algorithm is a commonly adopted approach in statistical exploratory analyses as well as in data mining. The algorithms applied most frequently to such type of research include the Lloyd, the MacQueen and the Hartigan and Wong algorithms (Everitt et al., 2011). These algorithms are relatively easy to use, have a large calculating potential and require relatively little computer memory compared to other clustering

algorithms. Research studies do not identify the best cluster analysis algorithm. The choice of a specific algorithm depends on the structure of a data set, its size, the number of analysed variables, etc. Due to large sizes of our data sets, we employed the Lloyd algorithm (implemented in the Statistica software). It is one of the most commonly used data mining algorithms. Its popularity stems from three main reasons (Lloyd, 1982):

- Minimizing an objective function is relatively easy and intuitive,
- The algorithm is simple, effective and often leads to optimal solutions,
- The results of the analysis are easily interpretable.

The characteristic feature of the methods for optimizing the initial partition of objects is an a priori determination of the number of clusters. One of the ways to conduct an analysis in this area is to estimate this number by means of the classification quality measures. However, as emphasized by Everitt et al. (2011), the selection of the optimal number of clusters should be done on the basis of the synthesis of the results obtained with the help of other methods. Such a procedure is recommended e.g. due to the fact that each method is based on predefined assumptions referring to the structure of classes, which are not always satisfied. Therefore, in our analysis, we applied several measures that are frequently implemented in empirical research studies and are available in the R package clusterSim: the Calinski-Harabasz index, the Krzanowski-Lai index, the Davies and Bouldin index, the Gap Statistic (Walesiak and Dudek, 2011). The hybridization procedure consists of the following steps:

1. The indication of the qualitative dependent variable and the set of independent variables within the data set,
2. The selection of quantitative independent variables from the set of independent variables and their application to building clusters, and subsequently the replacement of all the quantitative variables in the predictive model by the new variable informing about cluster membership,
   a. Subjective determination of the number of clusters or determination of the number by means of any cluster-validity measure,
   b. The reduction of the number of quantitative independent variables using the Random Forest if the number of such variables exceeds 15; more specifically, the selection of 15 variables on the basis of the variable-importance ranking,
3. The construction of a decision tree model by means of all qualitative independent variables and the new qualitative independent variable created in step 2.

We sequentially combined both analytical tools, thus creating a hybrid CART – k-means model. In the first step of the procedure, we created clusters on the basis of quantitative independent variables from the data set. The number of clusters could not exceed 15 (Blattberg et al., 2008). If a data set consisted of a larger number of quantitative variables, it was necessary to select 15 of them. This selection was carried out with the help of the Random Forest, which is

a method that allows creating a variable-importance ranking. The 15 variables thus selected were those most strongly related to the dependent variable. In the second step, a decision tree model was built, which comprised of qualitative independent variables and the new variable providing information about the cluster membership. The original quantitative variables were not used in the analysis.

The cluster analysis determined 2–20 clusters, which implied that the analysis of each data set yielded 19 different hybrid models. Setting the maximum number of clusters to 20 was our subjective choice. This value was higher than the maximum number of clusters indicated by the cluster validity measures used in the study. All quantitative variables used in the cluster analysis were standardized using the z-score formula ((value-mean)/standard deviation). We also calculated cluster validity measures, but their values did not determine the optimal number of clusters.

Additionally, a decision tree model based on the entire non-transformed set of independent variables (both categorical and numerical) was built for each data set (the so-called base tree). The decision tree is characterised by the following parameters: split rule – Gini measure, equal misclassifications costs, equal a priori probabilities, minimal number of cases in a parent node (5% of training set), minimal number of cases in a leaf (2% of training set) and maximum depth of the tree (15 levels). Its performance was a reference point for comparable hybrid models. The number of predictive models used for the purposes of the analysis totalled 280.

2.2 DESCRIPTION OF DATA SETS

Most of the data sets used in the experiment come from a well-known UCI machine learning repository (Asuncion and Newman, 2007). Table 1 provides information on the name of the data set, the number and type of independent variables, the number of categories of the dependent variable and the number of cases. Originally, this repository was intended to select data sets relating to the analytical CRM, database marketing and other business analytical areas. It was also important that the dependent variable was binary. Unfortunately, during the collection of data, it turned out that this type of data is confidential and is very rarely available in publicly-accessible online repositories. Ultimately, we decided to choose data sets with a varying number of cases (from 208 to 50,000), different numbers of dependent variable categories (from 2 to 10) and different numbers and types of independent variables (from 4 to 111). According to our intentions, this diversity was to ensure more reliable testing of hybrid models.

Each data set was divided into a training set (70 %), and a test set (30 %). The variables for which the missing data exceeded 10 percent, and the instances for which the missing data exceeded 50 percent, were excluded from the analysis. In the remaining cases, the missing data were substituted for by mean or modal values. We decided to replace the missing data by the simplest meth-

ods to eliminate their possible impact on the quality of the predictive models. The variables possessing unique values (ID, phone number, dates) were not analysed.

The models were assessed on the basis of measures calculated with the use of the misclassification matrix: accuracy ((TP + TN) / (TP + FP +TN + FN)), recall (TP / (TP + FN)), precision (TP / (TP + FP)) and F-measure ((2 × precision × recall) / (precision + recall)). The acronyms used in the formulas come from the confusion matrix and represent true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Additionally, the lift measure in the first and second decile of test set was calculated. The lift is the ratio of the response rate in a decile to the average response rate (in the whole data set).

TABLE 1. THE CHARACTERISTICS OF DATA SETS

Data set	Number of independent variables	Number of categories of dependent variable	Number of cases
(D1) Bank Marketing	6 numerical and 8 categorical	2	45211
(D2) German Credit	7 numerical and 13 categorical	2	1000
(D3) Insurance Company	23 numerical and 62 categorical	2	5822
(D4) Churn	15 numerical and 5 categorical	2	5000
(D5) KDD 2009 (preprocessed)	3 numerical and 18 categorical	2	50000
(D6) CINA Marketing	3 numerical and 108 categorical	2	16033
(D7) Australian Credit	6 numerical and 8 categorical	2	690
(D8) Banknote	4 numerical	2	1372
(D9) Heart (Statlog)	5 numerical and 8 categorical	2	270
(D10) Ionosphere	34 numerical	2	352
(D11) Pendigits	16 numerical	10	10992
(D12) Image Segment	14 numerical and 4 categorical	7	2310
(D13) Sonar	60 numerical	2	208
(D14) Vehicle	18 numerical	4	846

S o u r c e: own compilation.

3. RESULTS OF EXPERIMENT

Table 2 presents the selected performance measures for all data sets. The measure for the best hybrid model is placed in front of the bracket, whereas the measure for the base tree inside the bracket. In some cases, the difference was

in the third decimal place, but it is not visible after rounding the results. Data presented in the table indicate that in 8 out of 14 data sets (D1, D2, D6-D9, D11, and D13) a hybrid approach was more effective than unmodified decision tree, considering all the measures.

TABLE 2. INCREASED VALUES OF PERFORMANCE MEASURES IN HYBRID MODELS AS COMPARED WITH THE BASE TREE

Data set	Accuracy	Precision	Recall	*F*-measure	Lift 10%	Lift 20%
D1	0.87 (0.77)	0.45 (0.29)	0.78 (0.71)	0.46 (0.42)	4.23 (3.59)	3.20 (2.55)
D2	0.67 (0.66)	0.50 (0.49)	0.84 (0.77)	0.61 (0.60)	1.95 (1.50)	1.76 (1.50)
D3	0.71 (0.59)	0.14 (0.10)	0.75 (0.71)	0.22 (0.18)	3.17 (3.17)	2.51 (2.51)
D4	0.82 (0.86)	0.40 (0.49)	0.77 (0.78)	0.50 (0.60)	3.94 (4.21)	2.94 (3.59)
D5	0.69 (0.65)	0.10 (0.10)	0.73 (0.44)	0.16 (0.16)	1.85 (1.85)	1.45 (1.39)
D6	0.92 (0.90)	0.82 (0.76)	0.90 (0.87)	0.84 (0.81)	3.62 (3.11)	3.62 (3.11)
D7	0.89 (0.87)	0.90 (0.86)	0.91 (0.85)	0.88 (0.85)	2.17 (2.07)	2.09 (2.07)
D8	0.98 (0.91)	0.96 (0.90)	1.00 (0.92)	0.98 (0.91)	2.08 (2.06)	2.08 (2.06)
D9	0.82 (0.70)	0.77 (0.60)	0.80 (0.79)	0.77 (0.68)	2.23 (2.03)	2.10 (2.03)
D10	0.88 (0.88)	1.00 (1.00)	0.78 (0.64)	0.80 (0.78)	2.92 (2.92)	2.92 (2.92)
D11	0.85 (0.80)	0.97 (0.94)	0.98 (0.81)	0.96 (0.87)	8.84 (7.25)	4.83 (3.47)
D12	0.92 (0.92)	1.00 (0.96)	1.00 (1.00)	0.99 (0.98)	7.07 (6.79)	4.97 (1.00)
D13	0.85 (0.74)	0.89 (0.71)	0.93 (0.80)	0.85 (0.75)	2.07 (1.03)	2.07 (1.45)
D14	0.63 (0.46)	0.85 (0.85)	0.98 (0.84)	0.85 (0.85)	3.63 (3.57)	3.50 (3.50)

S o u r c e: own calculations.

Table 3 presents the minimal number of clusters which we needed to build the best hybrid model in our computer experiment. We intended to create the smallest possible number of clusters, which would facilitate their descriptions. Unfortunately, approximately 60% of the models yielded 10 or more clusters. Moreover, it turned out that the optimal number of clusters indicated by cluster validity measures did not provide best solutions. Also, when the number of clusters reached 20, it became possible that a higher value of performance measures could have been obtained for a larger number of clusters.

TABLE 3. MINIMUM NUMBER OF CLUSTERS IN THE BEST HYBRID MODELS

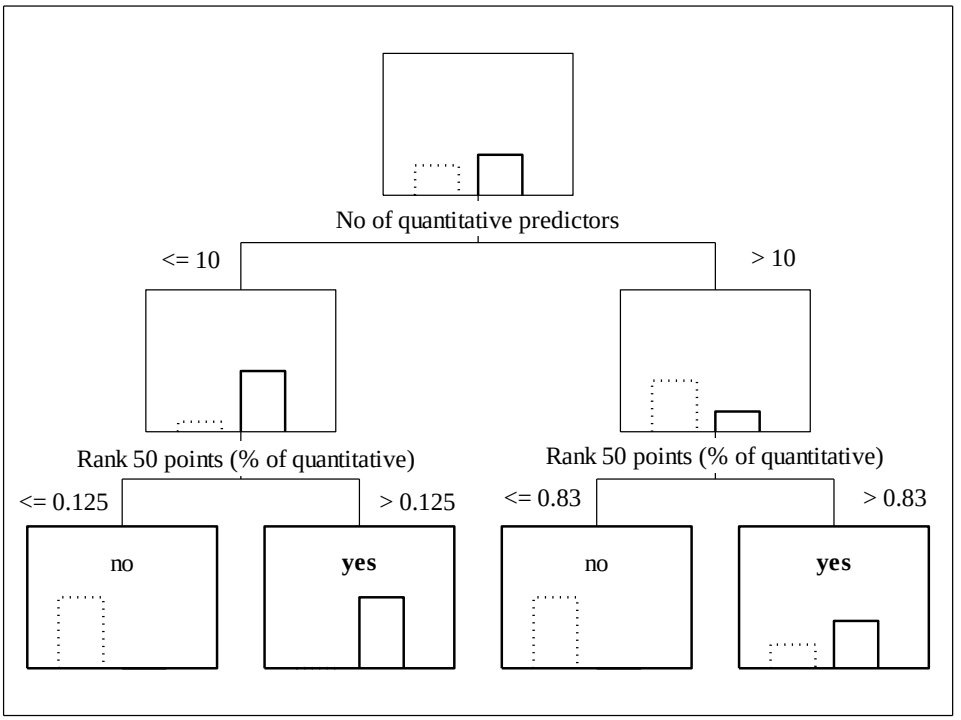
Data set	Accuracy	Precision	Recall	*F*-measure	Lift 10%	Lift 20%
D1	9	9	14	3	4	3
D2	3	12	10	16	17	9
D3	10	10	17	10	2	2
D4	7	7	14	17	8	14
D5	13	2	10	5	2	18
D6	20	20	2	20	20	20
D7	18	16	14	18	2	15
D8	20	20	13	20	11	11
D9	15	15	12	12	16	6
D10	13	18	9	13	10	18
D11	19	16	2	19	16	20
D12	18	3	4	4	2	3
D13	6	12	4	6	5	5
D14	19	16	2	18	19	18

S o u r c e: own compilation.

Subsequently, we investigated the reasons for the successes and failures of hybrid models. For this purpose, we employed the variable-importance ranking of the CART algorithm, which can assign from 0 to 100 points to all independent variables (Breiman et al., 1984). The higher the ranking position, the stronger the relationship between the predictor and the dependent variable.

In the next step we checked the relationship between the number of quantitative predictors with the largest number of assigned points and the quality of hybrid models. It was assumed that a strong relation between quantitative variables and the dependent variable indicates a strong relation between clusters and the dependent variable. Some other hypothetical success factors included the number of qualitative independent variables, the number of quantitative independent variables, the number of cases in the data set, the number of categories of the dependent variable and the difference in the numbers of observations among the categories of dependent variables (the latter variable provides information on the imbalance class problem). The next step involved building a decision tree in which the binary dependent variable assumed two values: 1 for the success of a hybrid model, and 0 for its failure. The set of independent variables comprised of all the above-mentioned determinants of the quality of hybrid models.

Figure 1. Decision tree classifying the result of the hybrid model
(binary dependent variable 'success' with two categories: yes / no)



S o u r c e: own compilation.

Figure 1 presents a CART decision tree model with 4 terminal nodes which encompass the best hybrid models (the leaves bear a 'yes' label). The highest quality of hybrid models was recorded for data sets where:

- the number of numerical independent variables was equal to or smaller than 10, and more than 12.5% of numerical predictors were assigned over 50 points (6 models),
- the number of numerical independent variables was larger than 10, and more than 83% of numerical predictors were assigned over 50 points (2 models).

In simplified terms, it can be stated that if a data set comprises 10 or fewer quantitative independent variables, hybrid models are more effective than a base tree. Such a result is obtained for 6 out of 7 sets. The success of a hybrid approach may result from the manner of dividing the classification tree. When quantitative predictors are used, the number of possible splits of nodes is equal to or smaller than n, where n indicates the number of predictor values. In the case of qualitative predictors, the number of possible splits is much larger, amounting to $2^{n-1}-1$, where n indicates the number of predictor categories. A larger number of possible splits can lead to a greater number of possible, and sometimes better, solutions.

TABLE 4. ERROR RATES AND ERROR RATES AFTER 10-FOLD CV (IN BRACKETS) FOR THE BEST HYBRID MODELS ESTIMATED ON THE BASIS OF THE TRAINING SET

Data set	D1	D2	D6	D7	D8	D9	D11	D13
base tree	0.2272 (0.2273)	0.2400 (0.3756)	0.0964 (0.0977)	0.1304 (0.1808)	0.0698 (0.0908)	0.1005 (0.1771)	0.1972 (0.2172)	0.0616 (0.2803)
2 clusters	0.2817 (0.2733)	0.2871 (0.3658)	0.1209 (0.1213)	0.1180 (0.1542)	0.4303 (0.4359)	0.0899 (0.1486)	0.7969 (0.8015)	0.3630 (0.3561)
3 clusters	0.1875 (0.1889)	0.2743 (0.3669)	0.1209 (0.1213)	0.1180 (0.1542)	0.3292 (0.3575)	0.1005 (0.1829)	0.7011 (0.7026)	0.3014 (0.2955)
4 clusters	0.2050 (0.1846)	0.2943 (0.3903)	0.0997 (0.1003)	0.1139 (0.1610)	0.2781 (0.2736)	0.1058 (0.1676)	0.6106 (0.6120)	0.2329 (0.2348)
5 clusters	0.1425 (0.1570)	0.2971 (0.3793)	0.0943 (0.0948)	0.1180 (0.1497)	0.1719 (0.1725)	0.1005 (0.1977)	0.5537 (0.5539)	0.3014 (0.3030)
6 clusters	0.1411 (0.1375)	0.2929 (0.3756)	0.1209 (0.1164)	0.1097 (0.1545)	0.1937 (0.1986)	0.1058 (0.1839)	0.4574 (0.4594)	0.2055 (0.2576)
7 clusters	0.1299 (0.1294)	0.3000 (0.3683)	0.0831 (0.0871)	0.1284 (0.1549)	0.1833 (0.1884)	0.0847 (0.1552)	0.4029 (0.4053)	0.2123 (0.2955)
8 clusters	0.1406 (0.1597)	0.2714 (0.3756)	0.1011 (0.1022)	0.1180 (0.1629)	0.1552 (0.1703)	0.0952 (0.1609)	0.3253 (0.3283)	0.2192 (0.2424)
9 clusters	0.1249 (0.1287)	0.2871 (0.3699)	0.0981 (0.0991)	0.1201 (0.1606)	0.1687 (0.1646)	0.1058 (0.1607)	0.2629 (0.2660)	0.1644 (0.1667)
10 clusters	0.2324 (0.2441)	0.3400 (0.3962)	0.0835 (0.0839)	0.1139 (0.1640)	0.1208 (0.1215)	0.0794 (0.2326)	0.2302 (0.2325)	0.1781 (0.2045)
11 clusters	0.1584 (0.2514)	0.2743 (0.3558)	0.0950 (0.0964)	0.1200 (0.1558)	0.1156 (0.1283)	0.0952 (0.2081)	0.2798 (0.2814)	0.1781 (0.1818)
12 clusters	0.2263 (0.2367)	0.2843 (0.3443)	0.0793 (0.0864)	0.1284 (0.1587)	0.1053 (0.1056)	0.0741 (0.2289)	0.2031 (0.2052)	0.1849 (0.2045)
13 clusters	0.2056 (0.1965)	0.2643 (0.3742)	0.0793 (0.0836)	0.1180 (0.1490)	0.0531 (0.0579)	0.0794 (0.1786)	0.2085 (0.2104)	0.1781 (0.2348)
14 clusters	0.2605 (0.2413)	0.3257 (0.3903)	0.0850 (0.0856)	0.1325 (0.1603)	0.1146 (0.1169)	0.0847 (0.1697)	0.2972 (0.3015)	0.1986 (0.2424)
15 clusters	0.1724 (0.2004)	0.2857 (0.3956)	0.0809 (0.0815)	0.1014 (0.1621)	0.0260 (0.0284)	0.0741 (0.1905)	0.2250 (0.2240)	0.1849 (0.2061)
16 clusters	0.1642 (0.1697)	0.2971 (0.3664)	0.0812 (0.0858)	0.1284 (0.1746)	0.0448 (0.0477)	0.0952 (0.2651)	0.1860 (0.1869)	0.1644 (0.1985)
17 clusters	0.1965 (0.2028)	0.2686 (0.3836)	0.0830 (0.0835)	0.1180 (0.1575)	0.0323 (0.0318)	0.0847 (0.1863)	0.1469 (0.1471)	0.1712 (0.2308)
18 clusters	0.1787 (0.1755)	0.2442 (0.3506)	0.0796 (0.0862)	0.1221 (0.1713)	0.0615 (0.0670)	0.0847 (0.2048)	0.1457 (0.1449)	0.1712 (0.1756)
19 clusters	0.2134 (0.2285)	0.2786 (0.3831)	0.0790 (0.0796)	0.1118 (0.1475)	0.0437 (0.0465)	0.0899 (0.2073)	0.1522 (0.1522)	0.1918 (0.2424)
20 clusters	0.2005 (0.2080)	0.2571 (0.3642)	0.0781 (0.0785)	0.1201 (0.1558)	0.0177 (0.0193)	0.1164 (0.2201)	0.1448 (0.1465)	0.1438 (0.1742)

S o u r c e: own compilation.

Table 4 presents error rates and error rates after 10-fold cross validation (in brackets). The figures refer only to those data sets for which the hybrid models provided the best performance measures. Both error rates were estimated using the training set, because only that set contained variables informing about the class membership. The test set was used twice during the model evaluation. Firstly, the cluster's membership was predicted on the basis of quantitative variables. Subsequently we predicted the class of variable Y. The comparison of both values made it possible to assess the stability of the results, although in this case it was limited to the training set.

4. CONCLUSIONS

Building hybrid models which combine decision tree algorithms with cluster analysis can, in some cases, improve the performance of predictive models. Prior to starting analytical research, it may be worthwhile checking the relationships between independent variables and the dependent variable. This refers in particular to the number of quantitative predictors and their position in the variable-importance ranking. The process of building clusters cannot rely on cluster validity measures, because they indicate different numbers of clusters, and do not always guarantee good quality of hybrid models.

The weakness of this approach is reflected by a large number of clusters in the best hybrids. The average number of clusters in the hybrid predictive model providing the highest accuracy value was 14. For the remaining best-performance measures, the average number of clusters was: recall (9 clusters), precision (15 clusters), F-measure (14 clusters), and lift in both deciles (11 clusters). This had a negative impact on the possible interpretation of a model, making a hybrid approach similar to a black box, which we intended to avoid. Our intention was to build a model that would have higher predictive power and at the same time would not lose the properties of decision trees, i.e. would yield a set of easily interpretable "if ... then ..." rules.

Undoubtedly, the limitation of this analytical experiment was the lack of cross-validated error rates that would be estimated on the basis of the entire data set. This made it impossible to assess the stability of the results. Moreover, we are aware that our approach should have been compared with univariate optimal binning methods. This is a popular method for transforming quantitative variables into qualitative ones.

It should be noted that despite the double use of a test set (firstly, when objects were assigned to clusters, and again in the process of deployment the decision tree model), performance measures assumed higher values than in the base tree. Furthermore, a higher quality of hybrid models was achieved, despite the sensitivity of cluster analysis to outliers or the risk resulting from finding artefactual solution (lack of natural clusters in data). These promising results encourage further research in this area. They could be extended by

utilising a larger number of data sets or the employment of different decision tree algorithms (C4.5 or CHAID) or cluster analysis algorithms (Mac Queen's or Hartigan and Wong's).

REFERENCES

Asuncion A., Newman D., (2007), UCI machine learning repository, http://archive.ics.uci.edu.

Blattberg R., Kim B. D., Neslin S., (2008), *Database Marketing – Analyzing and Managing Customers, 1st ed.*, Springer, New York. DOI: 10.1007/978-0-387-72579-6.

Bose I., Chen X., (2009), Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn, *Journal of Organizational Computing and Electronic Commerce*, 19(2), 133–151, DOI: 10.1080/10919390902821291.

Breiman L., Friedman J., Olshen R., Stone C., (1984), *Classification and Regression Trees, 1st ed. Wadsworth statistics / probability series*, Wadsworth Publishing Company, Belmont, California.

Chu B. H., Tsai M. S., Ho C. S., (2007), Toward a Hybrid Data Mining Model for Customer Retention, *Knowledge-Based Systems*, 20(8), 703–718. DOI: 10.1016/j.knosys.2006.10.003.

Everitt B., Landau S., Leese M. D. S., (2011), *Cluster Analysis, 5th ed. Wiley Series in Probability and Statistics*, John Wiley & Sons, Chichester, West Sussex. DOI: 10.1002/9780470977811.

Ferraretti D., Lamma E., Gamberoni G., Febo M., Di Cuia R., (2011), Integrating Clustering and Classification Techniques: A Case Study for Reservoir Facies Prediction, [in:] Ryżko D., Gawrysik P., Rybiński H., Kryszkieiwcz M., *Emerging Intelligent Technologies in Industry*, Springer, Berlin, Heidelberg, 21–34. DOI: 10.1007/978-3-642-22732-5_3.

Gaddam S., Phoha V., Balagani K., (2007), K-means + ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means Clustering and ID3 Decision Tree Learning Methods, *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 345–354.DOI: 10.1109/TKDE.2007.44.

Khan D., Mohamudally N., (2011), An Integration of *k*-means and Decision Tree (ID3) Towards a More Efficient Data Mining Algorithm, *Journal of Computing*, 3(12), 76–82, https://sites.google.com/site/journalofcomputing/volume-3-issue-12-december-2011.

Łapczyński M., Jefmański B., (2013), Impact of Cluster Validity Measures on Performance of Hybrid Models Based on K-means and Decision Trees, [in:] Perner P., (ed.), *Advances in Data Mining*, Ibai Publishing, Fockendorf, 153–162.

Łapczyński M., Surma J., (2012), Hybrid Predictive Models for Optimizing Marketing Banner Ad Campaign in Online Social Network, [in:] Stahlbock R., (ed), *Proceedings of the 2012 International Conference on Data Mining (DMIN)*, CSREA Press, Las Vegas, Nevada, 140–146.

Li Y., Deng Z., Qian Q., Xu R., (2011), Churn Forecast Based on Two-step Classification in Security Industry, *Intelligent Information Management*, 3(4), 160–165. DOI: 10.4236/iim.2011.34019.

Lloyd S., (1982), Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137, Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/TIT.1982.1056489.

Shouman M., Turner T., Stocker R., (2012), Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients, [in:] Stahlbock R., (ed), *Proceedings of the 2012 International Conference on Data Mining (DMIN)*, CSREA Press, Las Vegas, Nevada, 24–30.

Walesiak M., Dudek A., (2011), clusterSim: Searching for Optimal Clustering Procedure for a Data Set, https://cran.r-project.org/web/packages/clusterSim. R package version 0.47–3.

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Report from the XXXVIII Conference on Multivariate Statistical Analysis

Aleksandra BASZCZYŃSKA[1]
Katarzyna BOLONEK-LASOŃ[2]

The 38th edition of the scientific conference on Multivariate Statistical Analysis 2019 (MSA 2019) was held on November 4–6, 2019 in Łódź, Poland. The conference was organized by the Department of Statistical Methods of the University of Łódź, the Institute of Statistics and Demography of the University of Łódź and the Committee on Statistics and Econometrics of the Polish Academy of Sciences, and was co-financed from the fund for popularization of scientific activities of the Minister of Science and Higher Education (in accordance with the agreement No 712/P-DUN/202019). The Organizing Committee was presided over by prof. Czesław Domański, while prof. Aleksandra Baszczyńska and Katarzyna Bolonek-Lasoń served as scientific secretaries of the conference.

Conference themes included multivariate distributions, statistical tests, nonparametric inference, factor analysis, cluster analysis, discrimination analysis, Bayesian methods, stochastic analysis and application of statistical methods in finance, economy, capital market and risk management.

The conference was attended by 72 participants from the following universities and research institutes: Adam Mickiewicz University in Poznań, Poznań University of Economics and Business, Wrocław University of Economics and Business, Cracow University of Economics, University of Gdańsk, University of Economics in Katowice, University of Łódź, Warsaw University of Technology, University of Szczecin, University of Information Technology and Management in Rzeszów, Warsaw School of Economics, Warsaw University of Life Sciences, The Pomeranian University in Słupsk, Poznań University of Life Sciences, University of Milan-Bicocca, University of Hagen, Statistics Poland, Statistical Office in Poznań and Statistical Office in Łódź. During the conference 42 papers were presented in 15 plenary and parallel sessions.

The opening addresses were delivered by prof. Czesław Domański, prof. Antoni Różalski, Rector of the University of Łódź, and prof. Michał Przybyliński, Vice Dean of the Faculty of Economics and Sociology of the University of Łódź.

[1] University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods, 41/43 Rewolucji 1905 r. St., 90-214 Łódź, Poland, ORCID: https://orcid.org/0000-0002-4477-2438.
[2] University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods, 41/43 Rewolucji 1905 r. St. , 90-214 Łódź, Poland, ORCID: https://orcid.org/0000-0003-3741-0313.

The sessions devoted to the memory of distinguished representatives of the statistical thought recalled Jakub Kazimierz Haur, who was portrayed by prof. Czesław Domański, Marcin Kromer, whose achievements were presented by prof. Jerzy T. Kowaleski, as well as more contemporary, eminent statisticians: prof. Krystyna Katulska, prof. Mirosław Krzysztofiak, prof. Józef Kolonko, prof. Stanisław Wydymus and prof. Michał Major.

The subsequent sessions were chaired by prof. Czesław Domański, prof. Marek Walesiak, prof. Iwona Markowicz, prof. Wojciech Gamrot, prof. Wojciech Zieliński, prof. Alina Jędrzejczak, prof. Grzgorz Kończak, prof. Tomasz Żądło, prof. Grażyna Dehnel, prof. Grażyna Trzpiot, prof. Andrzej Dudek, and prof. Andrzej Bąk.

The following papers were presented at the conference:

- Andrzej Bąk: *Methods of imputation of missing data using the R program on the example of the Local Data Bank*;
- Maciej Beręsewicz and Katarzyna Zadroga, *Estimation of the number of illegally residing foreigners in Poland in 2017–2018 using Bayesian non-linear mixed count regression models*;
- Michał Bernardelli: *Identification of turning points in time series from the cryptocurrency market*;
- Jacek Białek: *Chain drift problem in the CPI measurement based on scanner data*;
- Beata Bieszk-Stolorz: *Selected models of recurrent events in the assessment of the risk of re-registration in the labour office*;
- Katarzyna Budny, *Multivariate Chebyshev's inequality – some bounds on the probability of a random vector taking values in the Euclidean ball*;
- Second Bwanakare and Marek Cierpiał-Wolan: *Generalised Cross-Entropy Econometrics vs conflicting cross-border (Big) data sources. National accounts updating*;
- Grażyna Dehnel and Marek Walesiak: *An assessment of social cohesion of Poland's provinces based on classic and interval-valued data*;
- Anna Denkowska and Stanisław Wanat: *Linkages and systemic risk in the European insurances sector: Some new evidence based on dynamic spanning trees*;
- Czesław Domański: *Some remarks about normality tests based on characteristics of stochastic processes*;
- Józef Dziechciarz and Marta Dziechciarz-Duda: *Selected aspects of households' well-being measurement*;
- Wojciech Gamrot: *Skala Likerta i współczynnik regresji (The Likert scale and the slope of regression)*;
- Małgorzata Graczyk and Bronisław Ceranka: *Some remarks about highly D-efficient spring balance weighing designs*;

- Małgorzata Graczyk and Bronisław Ceranka: *New results regarding the construction method of D-optimal chemical balance weighing designs*;
- Francesca Greselin, Andrea Cappozzo and Thomas Brendan Murphy: *Advances in learning from contaminated datasets*;
- Wioletta Grzenda: *Bayesian multinomial logit models for disordered categories in the analysis of the situation of young people in the labour market in Poland*;
- Stanisław Jaworski: *Some remarks about estimation of Polish unemployment rate*;
- Alina Jędrzejczak and Kamila Trzcińska: *Application of the Zenga Distribution to the analysis of household income in Poland by socio-economic group*;
- Adam Juszczak: *Application of web-scrapping in inflation measurement*;
- Grzegorz Kończak: *On permutation multivariate extension of McNemar test*;
- Jerzy Korzeniewski: *Determining semantic relatedness of concepts – modifications proposals*;
- Małgorzata Krzciuk: *On EBLUP under some linear mixed model with correlated random effects*;
- Mirosław Krzyśko, Waldemar Wołyński, Waldemar Ratajczak and Anna Kierczyńska: *Kernel discriminant coordinates in the case of geographically weighted temporal-spatial data with variable selection*;
- Marta Małecka: *Asymptotic Properties of Duration-Based VaR Backtests*;
- Iwona Markowicz and Paweł Baran: *Divergences in intra-Community trade: the case of Poland*;
- Hans-Joachim Mittag: *A new virtual library containing interactive learning objects for statistics education*;
- Dominika Polko-Zając: *On permutation tests for comparing multidimensional populations*;
- Aneta Ptak-Chmielewska: *Application of multidimensional classification to prediction of SME*;
- Elżbieta Roszko-Wójtowicz and Maria M. Grzelak: *Innovation activities and competitiveness of manufacturing divisions in Poland in the years 2009–2017*;
- Dominik Sieradzki and Wojciech Zieliński: *Sample allocation in estimation of proportion in finite populations*;
- Tomasz Stachurski: *On methods of median inference based on an estimator of the distribution function*;
- Agnieszka Stanimir: *Multivariate statistical methods in the analysis of multiple responses questions*;
- Piotr Sulewski: *Recognizing distributions rather than goodness-of-fit testing*;
- Krzysztof Szymoniak-Książek: *Properties of nonparametric isotropy tests*;
- Grażyna Trzpiot: *Seniors in cities and senior friendly cities analysis for selected Polish cities*;
- Łukasz Wawrowski: *Impact of dependent variable transformation on poverty rate estimates in poviats*;
- Jacek Wesołowski: *Optimal sample allocation in stratified sampling schemes - linear algebra methods and algorithms*;

- Ewa Wycinka and Beata Jackowska: *Competing risks models in estimation of companies life time*;
- Janusz L. Wywiał and Grzegorz Sitek: *On variance of sample matrix eigenvalue*;
- Artur Zaborski: *Triads or tetrads? Comparison of incomplete methods for measuring similarity in preferences*;
- Łukasz Ziarko: *On the possibility of using association analysis to describe the behaviour of contractors in public tenders*;
- Tomasz Żądło: *On generalization of Quatember's bootstrap*.

The next MSA conference is scheduled for November 16–18, 2020 in Łódź, Poland.

More information about the conference is available at www.msa.uni.lodz.pl.