Control and Cybernetics
vol. **54** (2025) No. 4
pages: 487-519
DOI: 10.2478/candc-2025-0017

Cluster-wise modelling: some issues considered through the example of migrations at municipality level in Poland*

by

Jan W. Owsiński[1*], Jarosław Stańczak[1], Przemysław Śleszyński[2] and Rafał Wiśniewski[2]

[1] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warszawa, Poland
[2] Institute of Geography and Spatial Organization, Polish Academy of Sciences, Twarda 51/55, 00-818 Warszawa, Poland
*owsinski@ibspan.waw.pl

Abstract: An experiment is presented of cluster-wise modelling, related to modelling and analysis of internal migration flows in Poland at the municipality level (some 2500 units). The experiment was carried out within a project, in which it was assumed that migration flows linearly depend upon unemployment. This simple dependence was positively verified, and the associated model error maps, corresponding to the consecutive years over two decades (2003-2022) provide a very clear and telling spatial image.

Yet, the models obtained were statistically rather feeble. So, it was decided to experiment, in particular, with a set of analogous models, identified for subsets (clusters) of municipalities, employing a simile of classical k-means clustering procedure. Given the known dependence of the outcome from the k-means-like procedure upon the starting point, various initial configurations were considered. The results are exemplified in this paper, the problems appearing indicated, and some broader conclusions are drawn therefrom.

Keywords: cluster-wise modelling, clustering, k-means algorithm, objectivity, migrations, unemployment, municipality, Poland

*Submitted: October 2025; Accepted: December 2025.

1. Introduction: migration modelling project

1.1. The course of study

The work, presented in this paper, was done largely in the framework project, devoted to analysis of migrations in Poland at the municipality level* This project was composed of several modules, and the here considered issue of modelling was taken up in only one of them. In this stream of work, an attempt was made to establish a possibly realistic models of migration at the municipality level, based on relation between unemployment and net migration flow to / from a given municipality (close to 2500 units in Poland). The preliminary results from this attempt can be found in Owsiński et al. (2024), while the data set used is available at Stańczak et al. (2024).

The results, related to the identified models for the years 2003-2022 (meaning altogether 20 annual models) for the entire populations of municipalities, implied that the hypothesis of a relation between unemployment and migration definitely holds in substantive terms, even though the model has the basic linear form with just one independent variable (unemployment). However, the statistical quality of models obtained was rather feeble (correlation coefficient between model results and true values of net migration being at +0.3 to +0.4). The substantive correctness of the model, and hence of the relation hypothesized, could be seen in (i) the values of model coefficients (see the comments further on), (ii) their consistency over the years (no haphazard jumps, but rather smooth changes, which could be referred to the corresponding macroeconomic conditions and their shifts) and (iii) the highly telling spatial image, produced by the model errors, with well visible regularities, which were considered not only correct, but could also be explained in a satisfactory manner. This image, its example being shown in Fig. 1, makes clearly apparent the areas around bigger agglomerations, where the actual positive net migration exceeded the model-based values, and also, largely, but by no means exclusively, the rural peripheral areas, where, on the other hand, the net outmigration was also bigger than predicted.

In this situation, having models definitely substantively correct and valid, but displaying quite poor statistical characteristics, it seemed to be rather straightforward to look for only slightly more complex models that should secure much better statistical properties. Hence, several models, being variations over the basic one, all of them with just one additional element included, were tried out. These models included, in addition to unemployment, such independent variables as (i) weighted distance to the nearest urban centers of different levels (county, province,...), measured by car travel time, (ii) magnitude of the near-

*The research of P. Śleszyński and R. Wiśniewski was financed by the National Science Centre in Poland, grant no. 2021/41/B/HS4/02035, "Influence of internal migrations on socio-economic development of Poland after 1989".

est centers (population), in a quasi-gravity form, or (iii) a nonlinear (quadratic) term, still referring to unemployment. The thus constructed models displayed, of course, better statistical properties than the basic model, but the improvement achieved was by no means as could be expected from any really tangible factor.

As a kind of countercheck, the basic model was identified yet for two kinds of aggregates of municipalities, first for the FUAs (functional urban areas), as defined for Poland by a team of geographers (some 400 units, i.e. roughly six municipalities per FUA), and then for the administrative units of provinces in the shape, in which they functioned in Poland between the 1970s and 1990s (49 units, meaning, on average, 50 municipalities per one such unit). The main reason behind this check was the expectation that some spatial phenomena may get smoothed over the aggregates considered and thus less of haphazard variability would lead to statistically better results. Of course, in order for the "smoothing" hypothesis to hold, the units considered would have to fulfil definite socio-economic conditions. The outcome for FUAs was disappointing (again – correct models, but equally feeble statistical properties), while for the former provinces the statistics were significantly better. This result was, again, seen as confirming the basic hypothesis of the study, with, evidently, FUAs not having been defined in an appropriate manner.

Against this background, another approach was tried out, consisting in formation of municipality clusters, centered around respective models, which would possibly well reflect the differences in these models and, at the same time, the categorization of municipalities with respect to these models. The procedure applied imitated the classical k-means clustering algorithm. It turned out that this approach yielded astonishingly better statistical results, as this will also be shown further on, along with respective discussion and conclusions.

1.2. The modelling insight

Let us emphasize that the purpose of the entire study was not to "explain migration" in terms of some econometric and/or theoretical model, but rather to verify the hypothesis that a simple, straightforward and easily interpretable model may provide sufficient power in representing the phenomenon. In this perspective, unemployment could be seen not only as a true-to-life explanatory variable, but as a representative of a group of variables, which are expected to be significantly correlated and do not need to appear explicitly in the models. Such variables are, in particular, local GDP per capita, wages, or incomes, entrepreneurship indices, and we assume that unemployment represents them to a large extent, even if, definitely, being negatively correlated. One could have also considered other variables, which might be less correlated with unemployment, such as presence and magnitude of the urban center(s), age structure etc. This may lead, on the one hand, to quite complex model forms, and, on the other hand, may also lead to self-explanatory models, which would be of little use.

Definitely, namely, migration can be modelled using various forms, usually based on more than one independent variable. Some recent examples can be seen in Mbachu (2022) or Kanga et al. (2024). Such models are, actually, a reflection of convictions as to the truly multidimensional character of migrations, see, e.g., Docquier, Peri and Ruyssen (2014), Jennissen (2003), or, for much earlier sources, Lee (1966) or Piore (1979).

In the project, from which we report here, the following prerequisites were adopted: (1) the model ought to be possibly simple and simply interpretable; (2) it should be possibly robust, at least regarding its general form and characteristics; (3) in connection with the preceding precepts, the independent variable(s) ought to represent a possibly broad socio-economic context, serving as a proxy for a group of potential independent variables.

It was agreed that unemployment within a spatial unit may play the role of such an independent variable, representing quite a spectrum of place-related characteristics. Then, in order to possibly avoid the averaging and mutual influence effects, the municipal level data were considered most appropriate. We dispose of flow matrices between pairs of municipalities (bilateral) for the years spanning two decades, 2003-2022 (see Śleszyński, 2024). This is quite an imposing database. The respective matrices contain a lot of zeroes, of course, and the numbers are very highly differentiated, even if only away from zero (take a flow from a small rural municipality to another similar one, and the flow between two large cities, for instance). So, it was decided to use the net migration indices for individual municipalities, rather than the matrix data.

The question of interrelation between the phenomena of migration and unemployment has a rich scientific literature, and that in both directions (influence of unemployment on migration flows, and vice versa). Just some selected examples of somewhat more recent studies are: Basile, Girard and Mantuano (2012), Basile et al. (2019), Baumann, Svec and Sanzari (2015), Blanchflower and Shadforth (2009), Card (2001), Dustman et al. (2003), Espinosa and Diaz-Emparanza (2021), Fendel (2014), Granato et al. (2015), Günther (2000), Heid and Larch (2012), Herzog, Schlottman and Boehm (1993), Huynh and Vo (2023), Jofre-Monseny (2014), Kilic, Yucesan and Ozekicioglu (2019), Kondoh and Yabuuchi (2012), Kulkolkarn and Potipiti (2007), Lemos and Portes (2008), Meng and Zhang (2010), White (2015), or Zabel (2012). The plethora of studies and various data sets definitely indicates the existence of a connection, yet without any decisive and sufficiently general form of the relation.

2. Models of migration at municipality level

2.1. The basic model

The basic considered model of migration as depending upon unemployment has the following form:

$$m_{it} = a_{0t} + a_{1_t} u_{it} + \epsilon_{it} \tag{1}$$

where i corresponds to municipalities (around 2 500 units in Poland), and t is the index of the year, m_{it} is net migration (inflow minus outflow) for municipality i in year t, expressed as a fraction of total population, u_{it} is the unemployment level in the same municipality, expressed as a fraction of working age population, and ϵ_{it} is the error of the model for municipality i. We are looking for the values of coefficients a_{0t} and a_{1t} on the basis of data for all municipalities for a given year. We identified the consecutive models for the years 2003 - 2022, i.e. 20 annual models.

Notwithstanding the fact that the model proposed appears to be very simple, the hypothesis behind it turned out to find quite strong confirmation through the following aspects (which are illustrated further on by an excerpt from the model identification results, given in Table 1, and an example of the spatial pattern of model error categories for municipalities, provided in the map of Fig. 1):

i. in all twenty models (2003 through 2022) the positive values of a_{0t} and negative values of a_{1t} were obtained. meaning that for small unemployment levels there is net in-migration, decreasing as unemployment increases (as this can be very well seen in Table 1);

ii. the values of a_{0t} and a_{1t} were in each case such that m_{it} would become negative at roughly the middle of the span of values of u_{it};

iii. the changes of values of a_{0t} and a_{1t} over time during the two decades analysed were by no means random or chaotic, but formed a distinct evolution over time, which could be analysed in terms of, for instance, the macroeconomic factors, or the socio-economic situation; it can be seen in Table 1 that the evolution involved, in particular, increase in the downward slope steepness, meaning, in fact, stronger influence of unemployment on migration;

iv. the spatial image of the magnitude and sign of the model errors, from high positive to extreme negative errors, had, over the years, a very telling character, easily interpretable, with the highest positive errors (net in-migration higher than determined by the model) observed in (predominantly rural) municipalities surrounding larger urban centers, and the extreme negative errors (net outmigration higher than the model based one) mainly concentrated in rural peripheral areas, with, however, numerous exceptions, constituted by the smaller urban centers; this kind of image, with larger urban centers, where model errors are moderate, surrounded by (rural) municipalities, featuring high positive error values, and relatively peripheral areas showing mostly high negative model errors, finds a very clear interpretation and explanation.

Table 1. Examples of model coefficient values for selected years, along with error characteristics

Year	a_0	a_1	Model error		
			max	mean	sum/sum of absolute values
2003	0.00497	-0.0318	0.0724	0.0045	$1.093 \cdot 10^{-14}$ / 11.24
2006	0.00617	-0.0527	0.0741	0.0048	$-9.479 \cdot 10^{-15}$ / 11.83
2009	0.00580	-0.0579	0.0531	0.0043	$3.370 \cdot 10^{-14}$ / 10.77
2012	0.00534	-0.0541	0.0396	0.0040	$1.799 \cdot 10^{-14}$ / 9.94
2015	0.00398	-0.0618	0.0339	0.0038	$-5.390 \cdot 10^{-15}$ / 9.41
2018	0.00332	-0.0841	0.0717	0.0043	$-7.771 \cdot 10^{-15}$ / 10.72
2021	0.00375	-0.0860	0.0433	0.0046	$3.239 \cdot 10^{-15}$ / 11.39
2022	0.00386	-0.0855	0.0382	0.0045	$-1.161 \cdot 10^{-15}$ / 11.17

2.2. Other models and auxiliary computations

The potential impact on the model quality from the *endogeneity* phenomenon, known in econometrics, was considered to be nonexistent, as the values of the correlation coefficient between unemployment and error were practically zero (at the level causing numerical problems, namely well below 10^{-14}).

It must be admitted, however, that the values of the correlation coefficients between the model output and the actual values were not very convincing, usually somewhat below +0,4. For socio-economic models, with a high variety of objects (here: municipalities) and their conditioning, such a value of correlation may not be that bad, especially if the model correctness is confirmed in substantive terms, as indicated previously. In particular, the regularity of the spatial images obtained, like that in Fig. 1, but also other characteristics could suggest that it should be relatively easy to improve the statistical model performance by adding some features to the model identified, perhaps even quite simple ones. That is why attempts were undertaken of identifying somewhat more complex models, with the hope of achieving distinctly statistically better results. The attempts consisted in adding one element to the model (1), meaning that three, not two, coefficients were being identified. These additions included:

(i) Weighted distance to the nearest centers of three different levels (local, county and province), measured by car travel time:

$$m_{it} = a_{0t} + a_{1t}b_{it} + a_{2t}d_{it} + \varepsilon_{it} \tag{2}$$

where the coefficient a_{2t} is a measure of the impact of the distance index of a given municipality to the nearest urban centres, taking into account their ranks in the administrative and settlement hierarchy (Śleszyński, 2016). The value of this distance index is a weighted sum of the time distances related to travelling by car on public roads to the nearest local, county, and provincial centres (the weights assigned are 20%, 25%, and 55%, respectively). It is, therefore, a synthetic indicator. In this way, an aspect, directly related to space and transport infrastructure, representing, among other things, the peripherality of a given municipality, was introduced into the model.

(ii) Magnitude of the nearest centers (in a kind of gravity setting):

$$m_{it} = a_{0t} + a_{1t}b_{it} + a_{2t}/\delta_{it} + \varepsilon_{it} \tag{3}$$

with the difference from (2) being not only the placement of the synthetic time distance measure in the denominator, in accordance with the principles of gravity models, but also the modification of the distance measure, here denoted δ_{it}, in such a way that the weights of the nearby provincial centres depended on their size (population). As is well known, gravity models, applied to various phenomena in space, including, in particular, determining the catchment areas of centres for specific products or resources, as well as in modelling migration, are based on the Newtonian law of universal gravitation. In cases where we are dealing with only one 'mass' (i.e., one, rather than, as in the classical case, two objects), subjected to gravity on the part of the other masses in the wider system, this model can still be used, in an appropriately modified form. Our analysis was therefore concerned with determining, as in the case of models (2) and (4), to what extent the introduction of such an additional element can improve the quality of the model.

(iii) Nonlinear (quadratic) dependence upon unemployment:

$$m_{it} = a_{0t} + a_{1t}b_{it} + a_{2t}b_{it}^2 + \varepsilon_{it} \tag{4}$$

meaning that the coefficient a_{2t} was added, showing the effect on migration of squared unemployment rate. This was an attempt to capture the possible non-linear dependence of migration on unemployment.

The results for all these extended models were, of course, statistically a bit better than for the basic model, but did not bring any qualitative improvement, which would have shown that the additional factor is of importance in the explanation of migration processes.

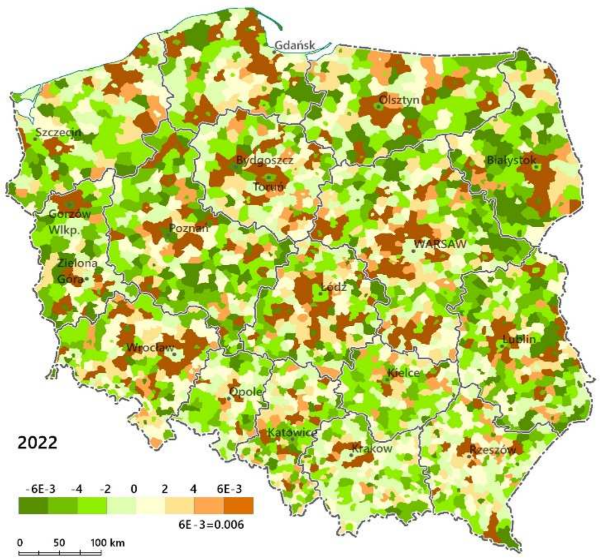

Figure 1. An example of error map of the basic linear model – the one for the year 2022

Exercises were also performed as to spatial units, constituting the aggregates of municipalities, first of all the functional urban areas (FUAs), more than 400 in Poland, and the previous provinces (49 in number), functioning for two decades at the end of the 20^{th} century. The results thereby obtained, though, as not directly related to the present subject, are not reported here.

Thus, in order to amend the quality of models at the municipality level, it was decided to determine the groups of municipalities, for which the basic model would be, in a respective version of it, most appropriate. For this purpose, a procedure, mimicking the classical k-means algorithm, was applied.

3. Obtaining group-wise migration models at municipality level

3.1. The procedure applied

The basic single-variable linear model of net migration for a municipality as dependent upon its unemployment level, proved to be correct, and yielded very interesting conclusions, but displayed weak correlation coefficients. So, a procedure was proposed, mimicking the well-known k-means algorithm of clustering, see, for instance, the generic paper by Steinhaus (1957) or the classical position from MacQueen (1967), who coined in the term "k-means". We use here the word "mimicking", for the designed and performed procedure, as presented below, is not equivalent to any of the formal versions of the k-means framework algorithm.

Thus, the proposed procedure worked as follows:

i. Basing on the results from the basic model (1), concerning all municipalities, we distinguish the following five groups of municipalities, namely those featuring: (I) high positive model error (more than 3/2 of standard deviation above zero); (II) medium positive error (between $\frac{1}{2}$ and 3/2 of standard deviation above zero); (III) small positive and negative error (between +1/2 and -1/2 of standard deviation with respect to model indications), i.e. the municipalities, for which the model turned out to be most "correct"; (IV) medium negative error (between $\frac{1}{2}$ and 3/2 of standard deviation below zero); and, finally, (V) high negative model error (more than 3/2 of standard deviation below zero).

ii. For the five groups (clusters) specified as before (i.e. in step i or in step iv), the respective five separate models in the form (1) are identified.

iii. For models from step ii, errors are calculated with respect to all (five) of them, for all municipalities (hence: five error values for each municipality).

iv. Municipalities are assigned to models determined on the basis of minimum error (a municipality is assigned the model, for which it displays the minimum error out of five). New groups (clusters) are thus established, composed of the municipalities with the smallest error with respect to a given model.

v. A stop criterion is checked, either a bound on the change of composition of clusters (e.g. less than 1% of membership changed?), or, simply (and most frequently) – a definite number of iterations achieved (e.g. 15). If the criterion is satisfied, the procedure STOPS, otherwise, - go to step ii.

3.2. Some important observations

It is a widely known feature of the k-means-like clustering procedure that the ultimate results, i.e. at fulfilment of a restrictive stop criterion, like "no change in the composition of clusters", should not differ much from those obtained through execution of a couple of iterations (fast convergence).

Since in this case the calculations are somewhat more complex than in the standard clustering problem, solved through application of classical k-means procedure, we limited the number of iterations to 10. It was hoped that we would obtain consistent groups of municipalities in the sense of the migration-unemployment model, corresponding to them. The composition of these groups (and their difference with respect to the starting group, in the first instance tried out based on initial error magnitude) and the corresponding models were of primary interest.

The procedure employed, "imitating" the standard k-means generic algorithm, should produce results near to "optimal", in the sense of approximating the minimum of the total sum of errors, for the given number of clusters and the given starting point (i.e. we should obtain, in each calculation, an approximation of a local solution). This dependence upon the number of clusters and the starting point is the property of the generic k-means algorithm. It is treated in practice, with the purpose of overcoming the problem of locality, in a variety of manners, but first of all by using multiple starting points and performing the entire procedure for each of them. Another way, of a similar nature, is to perform some introductory search for the possibly best starting point(s). This, of course, does not resolve the issue of dependence on the number of clusters.

And so, given the dependence upon the starting point of the procedure, a number of exercises were performed in order to verify the robustness of results, as this will be shown later on.

However, it must be emphasized that the procedure proposed diverges from the original k-means by the way, in which the assignment of objects (municipalities) to models is done. Namely, in this case, we deal with the models of the form

$$m_{it} = a_{0tq} + a_{1tq}u_{it} + \epsilon_{itq} \tag{5}$$

where index q refers to models for particular clusters (in the initial case, described before, $q = 1, 2, \ldots 5$). In formula (5) we assume that municipality i has been placed in model-related cluster q, in order to simplify the notation. Actually, instead of q we should write $q(i)$, meaning that the index of model, q, is the one that corresponds to the given municipality, i. (Further on we shall omit the annual index t as having lesser importance for our considerations.)

The classical k-means algorithm assigns objects to cluster prototypes by means of distance minimization. In our case this would amount to calculation, for each municipality i, of the following values:

$$(m_i - p_q(m_i))^2 + (u_i - p_q(u_i))^2 \tag{6}$$

for each q, with $p_q(m_i)$ denoting the value of coordinate m of the perpendicular projection of point (m_i, u_i) onto the line, representing model q, and analogously for $p_q(u_i)$, indicating the value of coordinate u. Formula (6) expresses the fact that in the classical k-means reference is made to squared Euclidean distance. It is for this distance measure that important properties, such as convergence, have been demonstrated for the k-means algorithm, when the clusters are represented by the respective means (the cluster mean minimizing the sum of distances to all elements of a cluster when distances are squared Euclidean). It is true that in computational practice, many other distance measures are being employed in this framework (most often – usual Euclidean distance). Yet, in our case, where we use model error:

$$m_i - (a_{0q} + a_{1q}u_i) = \epsilon_{iq} \tag{7}$$

the divergence from the distance-based concept of the algorithm is significant.

Since, however, we cannot analytically establish the consequences of this divergence, we are left with the simulation-like analysis, consisting in (sufficiently) numerous computational experiments. Actually, in our case, defined by the data set considered, the situation is even more difficult, as we shall see later on.

4. The initial results

We start this presentation from the exact performance of the procedure, described before in Section 3.1, including the way the starting points for consecutive years were obtained (i.e. on the basis of model errors).

The calculations were performed for particular years, available in the database, spanning two decades, 2003-2022. We start with the table of initial results, showing an exemplary starting point, based on the straight results for the basic linear model for the year 2022, meaning the outcome from steps i and ii of the procedure, previously outlined (breakdown into initial groups and calculation of respective models).

Already this table brings very interesting and relevant results, demonstrating that the attempt was, indeed, worthwhile. First, the correlation coefficients for the individual clusters are far higher than for the overall model, which could, of course, be anticipated. Yet, the scale of improvement is very much telling, especially if we consider the extreme groups, where outliers might (and certainly do) exert an important influence. Second, the partial models differ distinctly (see the values of a_{0q} and a_{1q} in subsequent rows of the table). Thus, we deal with a sequence of models, starting with the one, featuring the highest initial value (in-migration level for hypothetical zero unemployment) and the steepest decline along unemployment level, and ending with the one, where even for zero unemployment a negative net migration would occur, while the decline with rising unemployment is average. The share of municipalities, corresponding to the latter model, however, is marginal (some 2.5% of the total number). Two models with negative migration balance for zero unemployment were identified, but they distinctly differ, both in terms of the model coefficient values, and in the number of municipalities assigned.

The subsequent Table 3 shows the same kind of results, but for the initial year of the investigations, namely 2003. The observations, concerning the

Table 2. An example for the starting point of the k-means-like procedure (year 2022) based on model error: initial groups and corresponding models

Municipality groups according to model error	Number of units in a group	a_{0q}	a_{1q}	R – correlation coefficient	Model error		
					max	average	sum/sum of absolute values
High positive	188	0.0209	-0.1350	0.626	0.0218	0.0040	$1.086*10^{-15}/0.759$
Medium positive	378	0.0096	-0.0903	0.826	0.0039	0.0015	$9.593*10^{-16}/0.561$
Small	1126	0.0032	-0.0791	0.743	0.0036	0.0015	$-1.750*10^{-15}/1.643$
Medium negative	727	-0.0018	-0.0788	0.696	0.0039	0.0013	$-1.652*10^{-15}/0.959$
High negative	60	-0.0071	-0.0929	0.674	0.0072	0.0013	$-1.613*10^{-16}/0.078$

content of Table 1, remain certainly largely valid for this example, as they are, actually, for virtually all of the twenty years considered. This concerns the qualitative description of the results, while the concrete numbers differ significantly, showing the evolution of the respective situation over two decades. An interesting aspect is constituted by the clearly lower values of the correlation coefficient, showing bigger dispersion of observations in this initial year of investigations.

For purposes of illustration, but also for the sake of comparison that shall be made later on, we show here the course of the lines, constituting the models, for the situation, described in Table 2.

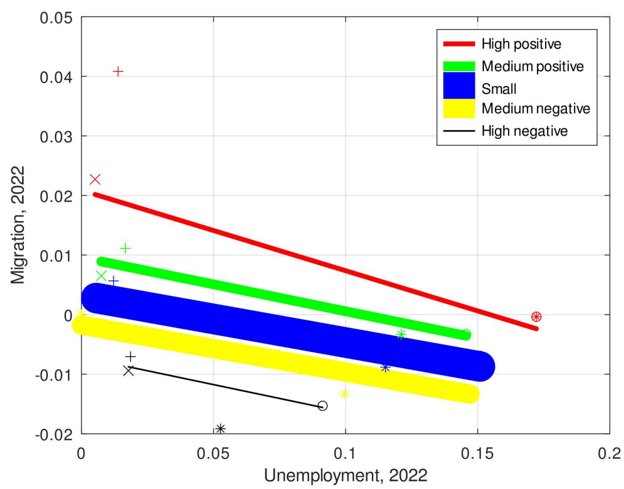

Figure 2. The course of models from Table 2. The stretch of lines is defined by the extreme observations, assigned to it. The figures show also characteristic points, associated with each model. Line thickness corresponds to group cardinality

Now, we shall pass to the results for step no. 10 in the same two years, 2022 and 2003, their starting points characterized in Tables 2 and 3, respectively, these results after 10 iterations being shown in Tables 4 and 5, respectively.

Here, we can see: (a) an important change in the numbers of municipalities, constituting individual "categories", (b) an essential improvement in the correlation coefficients, corresponding to individual municipality types identified (especially so as regards the initial year of the study, 2003), and (c) the consistent preservation of the character of the five distinct models, with even more intuitive expression for the last of these (not only the lowest – negative –

Table 3. Example for the starting point of the k-means-like procedure (year 2003): the initial groups of municipalities and the corresponding models

Municipality groups according to model error	Group cardinality	a_{0q}	a_{1q}	R - correlation coefficient	Model error		
					max	average	sum/sum of absolute values
High positive	156	0.0276	-0.0612	0.407	0.0524	0.00659	$1.433 * 10^{-15}/1.017$
Medium positive	322	0.0114	-0.0346	0.754	0.0043	0.0016	$-3.025 * 10^{-15}/0.516$
Small	1310	0.0042	-0.0029	0.670	0.0039	0.00154	$-2.271 * 10^{-15}/2.021$
Medium negative	644	-0.0013	-0.0286	0.624	0.0045	0.0015	$9.116 * 10^{-16}/0.982$
High negative	47	-0.0055	-0.0447	0.675	0.0166	0.0017	$4.857 * 10^{-17}/0.079$

Table 4. Results after 10 iterations for the case of Table 2, year 2022

Municipalities according to initial model error	Number of units	a_{0q}	a_{1q}	R – correlation coefficient	Model error		
					max	average	sum/sum of absolute values
High positive	135	0.0237	-0.1560	0.770	0.0193	0.0038	$-9.89 * 10^{-17}/0.514$
Medium positive	414	0.0117	-0.1172	0.839	0.0055	0.0018	-1.30*10^{-15}/0.757
Small	823	0.0045	-0.0864	0.825	0.0031	0.0012	1.31*10^{-15}/0.970
Medium negative	818	-0.0005	-0.0687	0.763	0.0023	0.0010	-1.87*10^{-15}/0.844
High negative	289	-0.0051	-0.0670	0.593	0.0106	0.0015	1.07*10^{-15}/0.424

Table 5. Results after 10 iterations for the case of Table 3, year 2003

Municipalities according to initial model error	Number of units	a_{0q}	a_{1q}	R – correlation coefficient	Model error		
					max	average	sum/sum of absolute values
High positive	78	0.0367	-0.0809	0.595	0.0450	0.0062	$1.99*10^{-16}/0.48$
Medium positive	366	0.0172	-0.0590	0.808	0.0086	0.0023	$-3.26*10^{-15}/0.86$
Small	918	0.0070	-0.0394	0.814	0.0041	0.0013	$-4.09*10^{-15}/1.20$
Medium negative	839	-0.0016	-0.0329	0.795	0.0027	0.0011	$-4.59*10^{-15}/0.94$
High negative	278	-0.0032	-0.0356	0.621	0.0212	0.0015	$-1.56*10^{-16}/0.43$

coefficient a_0, but also less negative slope of the line, reflecting the dependence on unemployment level).

We shall return to these aspects, especially the third one, in the results, shown and discussed in Section 6 of the paper.

In analogy to Fig. 2, we show in Fig. 3 the coure of the models for step 10 in the year 2022.

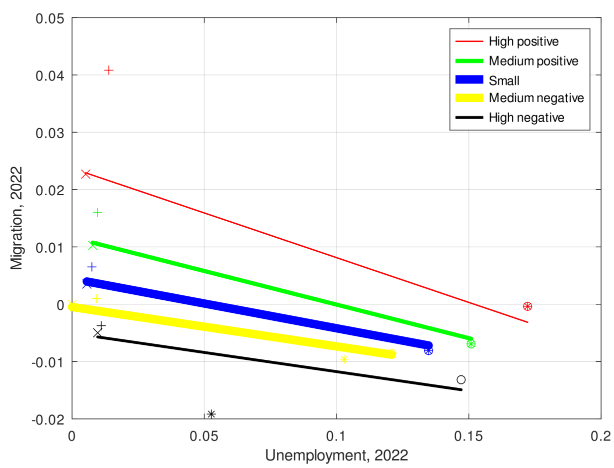

Figure 3. The course of models from Table 4. The stretch of lines is defined by the extreme observations, assigned to it. The figures show also characteristic points, associated with each model. Line thickness corresponds to group cardinality

5. The spatial aspect

Already the changes in the proportions of groups of municipalities in the entire population, taking place along the procedure, are very telling with respect to the character of units, forming these groups. A visual assessment can easily be done by comparing Fig. 1, showing an example of the initial classification of municipalities, according to the error of the single model, with Fig. 4, where a partly analogous image is shown, namely that of municipality categories after 10 iterations of the procedure for two selected years. Hence, we deal no longer with the “error-value-classes” (as in Fig. 1), but with municipality categories, corresponding to different models, obtained via the procedure, which started from exactly such “error-value-classes” (as shown here in Fig. 4).

The differences in the spatial images of the two figures are clear, but so are, as well, those between the two cases, shown in Fig. 4, coming from the beginning and the end of the period considered (2003-2022). It can be seen that the spatial image of Poland, in the domain of migration vs. unemployment, observed here in relative terms, i.e. with respect to five models, corresponding to subgroups of municipalities, has, indeed, changed. The essence of the change consists in, very roughly, (1) the fact that Eastern Poland, especially the Northeastern Poland, ceased to be the very pronounced outmigration area against the background of Poland at large, and (2) an increase in the migration-related differences in space (more of stronger high in-migration foci of local and regional scale, as compared to the year 2003).

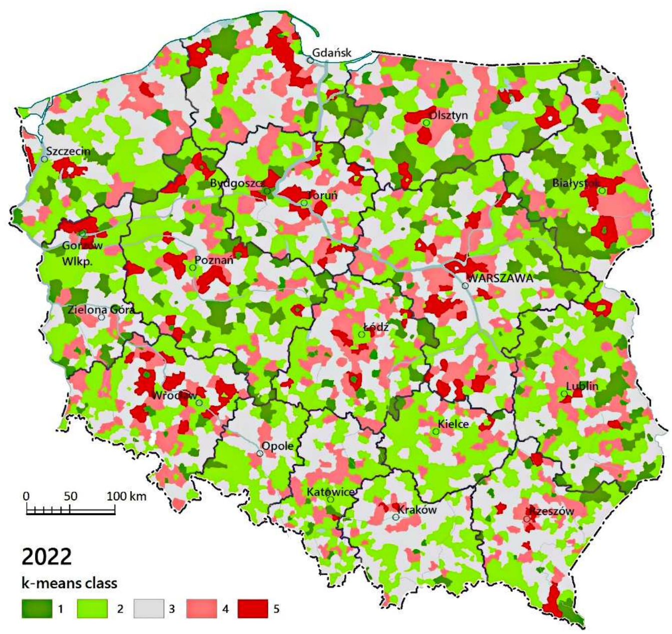

Figure 4. Two examples of the spatial image of municipality classes, obtained through the clustering procedure, for the years 2003 and 2022, the initial and the final years of the study

Yet, while these results appear – again, like with the basic model for all municipalities – to be satisfactory, not only confirming definite intuitions or knowledge from other sources, there is a doubt as to the essential nature, validity and objectivity of the results, even if they definitely are of significant value. That is why a series of further experiments was designed, primarily based on selecting different starting points, and observing more closely the evolution of the intermediary solutions, reported at length in Section 6.

6. The study of the influence of the starting points

6.1. The basic facts and assumptions

It is quite obvious that we deal here with a relatively homogeneous distribution of observations in the assumed two-dimensional space (migration-unemployment), where there are no "objective" boundaries between any hypothetical subgroups, be it due to some "external" factors, like, e.g., differences in formal types of municipalities (*rural, urban, urban-rural* being the three formal categories, distinguished in Poland) or some other variables, not accounted for here (e.g. scale of population in the municipality – from large cities down to sparsely populated peripheral municipalities).

In view of this, it can be expected that also the model-related clusters do not constitute well-defined and separated subgroups, and it is only the optimizing procedure that determines the "areas of assignment to a model" in the space considered.

This is closely associated with the properties of the k-means-like procedure, mentioned before, namely dependence on the number of clusters and on the starting point. We do not analyze the first of these, but concentrate on the second. Actually, in conditions of relatively homogeneous set of observations, this dependence may get quite significant, with a rather flat landscape of the objective function, featuring only minor, but very dense, local variations.

The experiments, whose three examples shall be treated here, consisted, therefore, of starting the very same procedure from different starting points, different also from that used in the initial study, described before, when the starting point was determined on the basis of model errors. In addition, the experiments performed and here analysed used, in each case, not five, but six models ($q = 1, 2, \cdots, 6$).

6.2. Experiment no. 1

Figure 5 shows the cloud of data for municipalities for the year 2003 in the space of migration vs. unemployment and, against this background, the assumed six models, constituting the starting point.

The initial models, shown in Fig. 5, are characterized in Table 6. There is, obviously, a very clear difference with respect to the basic case considered before, founded upon the model error values. The numbering of models starts from the one with the highest a_0 value, corresponding to the lowest a_1. Figure 5 illustrates also the fact of relative homogeneity of the data cloud.

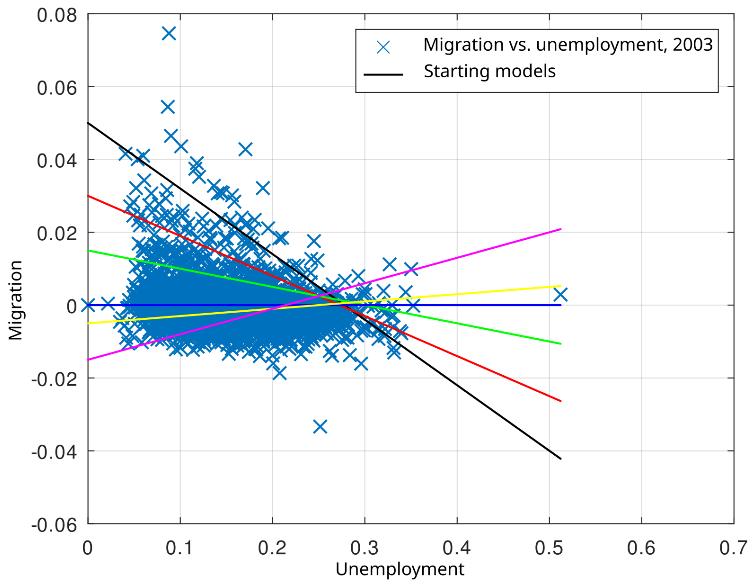

Figure 5. The data plot for the year 2003 and the initial models of dependence of migration upon unemployment

Table 6. Characteristics of the starting point models in the exercise no. 1

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.05	0.03	0.015	0.0	-0.005	-0.015
a_{1q}	-0.18	-0.11	-0.05	0.0	0.02	0.07

The situation after the first step of the procedure is shown in Fig. 6. This figure (and the next ones) shows the division of the set of municipalities into clusters from the preceding step of the procedure and the models, identified for this division (and not vice versa, i.e. the models and the municipalities assigned to them). This fact, along with the use of error for assignment, rather than distance, causes some intuitively strange phenomena, which can be observed in the figures. The respective characteristics are provided in Table 7.

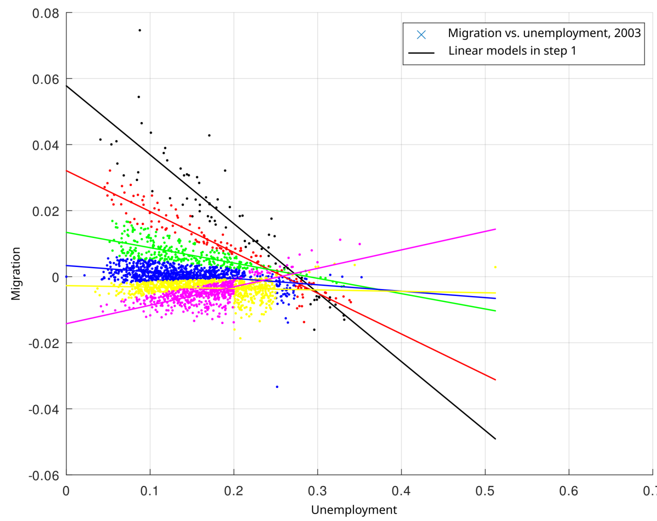

Figure 6. The division of the set of municipalities on the basis of initial models and the models, obtained for this division, exercise no. 1, step 1

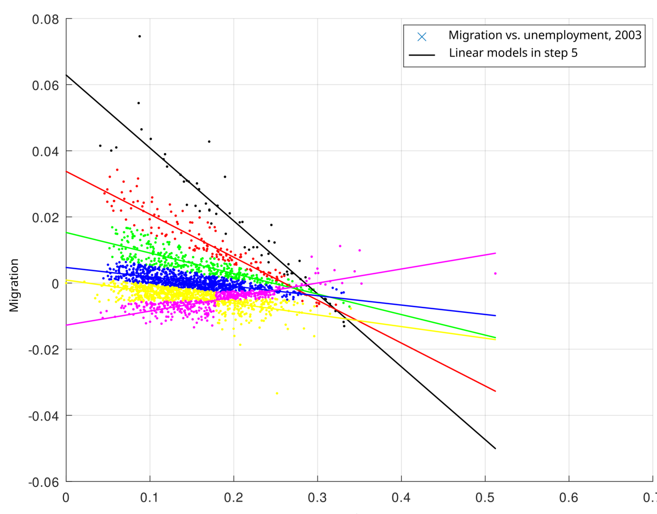

Figure 7. The division of the set of municipalities on the basis of models from step 4 and the models, obtained for this division, exercise no. 1, step 5

Table 7. Characteristics of models after step 1 in the exercise no. 1

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.058	0.033	0.013	0.003	-0.003	-0.014
a_{1q}	-0.21	-0.12	-0.046	-0.02	-0.004	0.056
Number of municipalities assigned	73	126	284	804	670	522
Correlation coefficient	0.91	0.96	0.73	0.43	0.10	0.67

For illustration, let us yet observe the situation at step 5 of the procedure, as shown in Table 8 and then in Fig. 7.

Table 8. Characteristics of models after step 5 in the exercise no. 1

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.063	0.034	0.015	0.007	0.001	-0.013
a_{1q}	-0.22	-0.13	-0.06	-0.03	-0.035	0.04
Number of municipalities assigned	57	147	333	855	677	410
Correlation coefficient	0.93	0.95	0.86	0.70	0.68	0.79

Rather than drawing any conclusions on the basis of the already presented results, we shall continue with presentation of the experiments.

6.3. Experiment no. 2

Like before, we begin with the presentation of the initial situation for this experiment, see Fig. 8 and Table 9.

Thus, we consider the same data set as before (that is, the one for the year 2003), but the initial models, characterized in Fig. 8 and Table 9, differ significantly from those, assumed for the experiment no. 1. In order not to use too much space, we shall present now only the results, obtained after step 5 of the applied procedure. They are provided in Fig. 9 and Table 10.

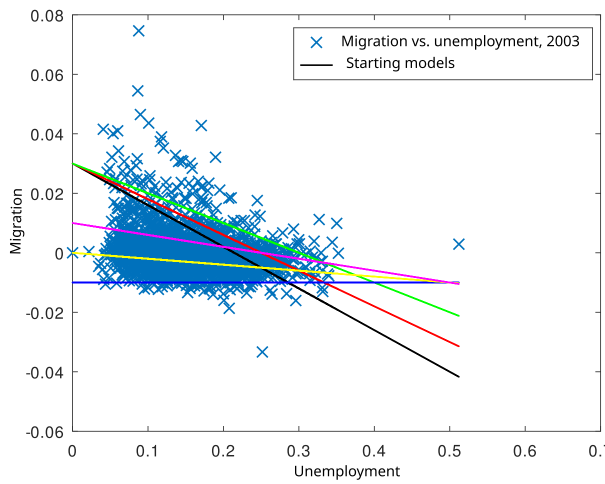

Figure 8. The data plot for the year 2003 and the initial models of dependence of migration upon unemployment (experiment no. 2)

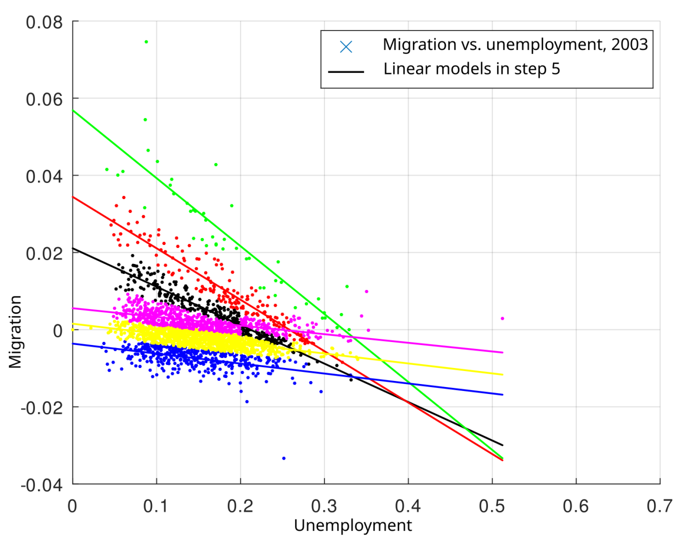

Figure 9. The division of the set of municipalities on the basis of models from step 4 and the models, obtained for this division, exercise no. 2, step 5

Table 9. Characteristics of the starting point models in the experiment no. 2

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.03	0.03	0.03	-0.01	0.00	0.01
a_{1q}	-0.14	-0.12	-0.10	0.00	-0.02	-0.04

Table 10. Characteristics of models after step 5 in the exercise no. 2

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.021	0.034	0.057	-0.004	0.002	0.006
a_{1q}	-0.100	-0.133	-0.176	-0.026	-0.026	-0.022
Number of municipalities assigned	290	172	46	395	995	581
Correlation coefficient	0.96	0.95	0.88	0.49	0.70	0.63

6.4. Experiment no. 3

6.4.1. Generation of the starting point

This experiment had a somewhat different character than the two, described before. Namely, the starting point was not just given, but was selected with the use of a special procedure for generating a possibly good, if not "the best" starting point. This heuristic procedure consisted in generation of potential starting points and evaluating them for their quality. Actually, two kinds of such procedure were tried out. We present here the results for one of them. Like before, figures and tables illustrate the initial situation and the one after the iteration 5 of the procedure.

The heuristic procedures, applied to obtain good starting points for the k-means-like clustering algorithm were developed so as to optimize the initial models by finding the solution minimizing the sum of absolute values of component models' errors in the assumed step of searching for a model (the model obtained from the k-means-like clustering algorithm in an assumed step of the proper procedure).

First, a very simple heuristic method, called the *fastest growth algorithm* was proposed, in order to check if the device really works:

1. Randomly generate the starting point (i.e. starting model coefficients for the k-means-like clustering procedure).
2. Compute the quality of the first set of initial models (the sum of absolute values of component models' errors in the assumed step of searching for a model).
3. This solution becomes a base solution.
4. Randomly modify parameters of several copies of the entire set of models considered (5 in the calculations actually performed) of the base solution.
5. Compute the quality of the obtained model copies.
6. Select the best one.
7. Compare it with the base solution, if it is better, replace the base solution with it, if not – the base solution remains unchanged.
8. If the stop condition is fulfilled, then return the best solution found, otherwise go to point 4.

The stop condition was constituted by the number of iterations, here: 1 000.

Since the use of such approach turned out to be computationally and substantively feasible, we used a much better heuristic, based on an evolutionary algorithm:

1. Randomly generate the population (30) of starting points, that is – 30 complete model sets.
2. Compute the quality of the individuals in the population of model sets.
3. Perform reproduction and crossover of parent (sub)population to obtain the offspring (sub)population (the size is 5 x the parent population).
4. Perform mutation of offspring (sub)popu-lation.
5. Compute the quality of individuals in the obtained offspring (sub)population.
6. Select new parent population using tournament selection.
7. If the stop condition is fulfilled, then return the best solution, otherwise go to point 3.

The stop condition is again constituted by the number of iterations, 100 in this case. The used crossover method is an averaging crossover, the mutation is a standard method with probability of execution 0.1 and normal distribution. The second method, as expected, produces somewhat better solutions.

6.4.2. The results for Experiment no. 3

Like for the previous experiments (that is: no. 1, no. 2 and the initial one, based on model error magnitude), we present here analogous kinds of results, beginning with the characteristics of one of the starting points obtained.

Table 11. Characteristics of the starting point models in the experiment no. 3

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.0037	-0.0061	0.0032	0.0053	0.0068	0.0091
a_{1q}	0.014	-0.012	-0.017	-0.050	-0.042	-0.011

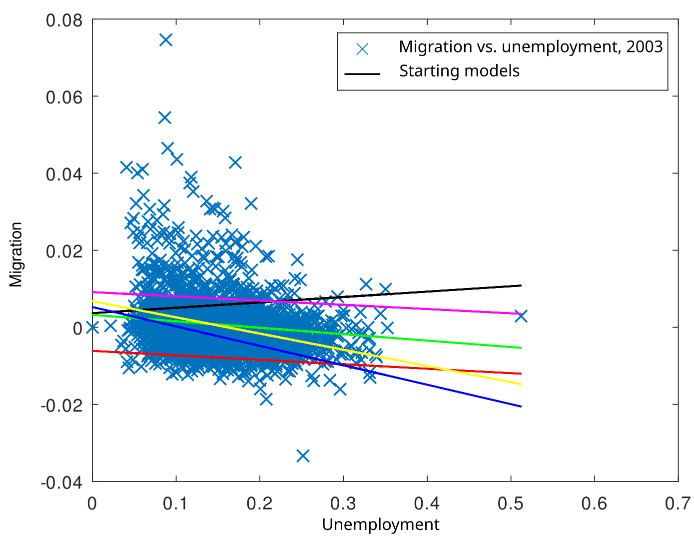

Figure 10. The data plot for the year 2003 and the initial models of dependence of migration upon unemployment (experiment no. 3)

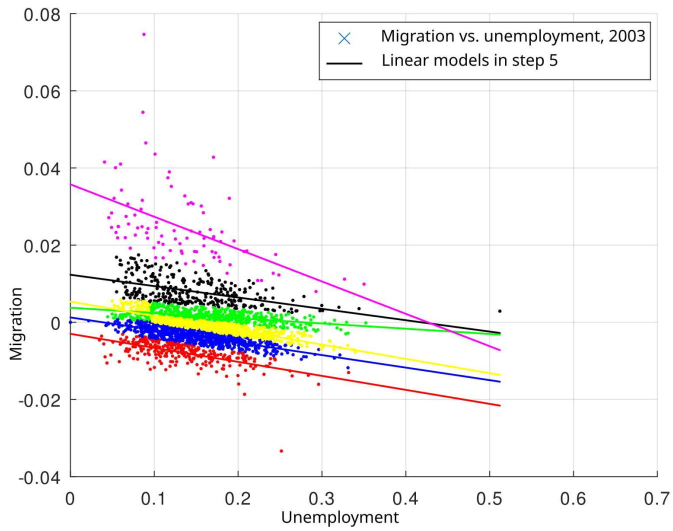

Figure 11. The data plot for the year 2003 and the models of dependence of migration upon unemployment in step 5 (experiment no. 3)

Table 12. Characteristics of models after step 5 in the exercise no. 3

Items	Model number					
	1	2	3	4	5	6
a_{0q}	0.012	-0.003	0.004	0.001	0.005	0.036
a_{1q}	-0.029	-0.036	-0.014	-0.033	-0.037	-0.084
Number of municipalities assigned	313	287	485	687	619	88
Correlation coefficient	0.52	0.63	0.56	0.83	0.90	0.51

At this point we terminate the presentation of results, which shall be discussed in the next section.

7. Discussion

The overall purpose of the study was to obtain possibly well founded, in substantive and in statistical terms, partial models of migration vs. unemployment, so as to (1) improve the statistical quality over the model for the whole population of municipalities, and, possibly, (2) obtain partial models displaying sensible interpretative characteristics. Given the known dependence of the k-means-like procedures on the starting point, several starting points (model sets) were used. The results, obtained after a couple of iterations, along with the ones for the starting points, are provided here for three such starting points, apart from the initial study, in which a different number of clusters (municipality subsets) was assumed.

As already indicated, the obtained partial, cluster-wise models feature much better correlation characteristics than the global model, the fact that could, of course, be anticipated. Irrespective of the starting point, there is an evolution of the models, definitely towards the substantive "correctness", even if some models still correspond to marginal groups of municipalities in the sense of the model form assumed.

There are, however, several points of doubt, confirming the reservation, expressed at the end of Section 6.1, namely – whether we do not deal with a kind of artifacts rather than sound qualitative distinctions.

First, the evolution, mentioned above, performed through the k-means-like procedure, is much slower than anticipated. Indeed, even after 10 iterations the model sets still preserve the general initial character, with only quite moderate modifications of the overall shape. It is hard to say whether this is due to the specific features of this particular procedure (e.g. use of error instead of distance), or, rather, the nature of the data set (relatively homogeneous distribution of data points). In any case, even prolongation of the functioning of the procedure up to 30 iterations does not change much the image.

An illustration of the way this evolution takes place is provided in Fig. 12, where the values of coefficients of the models are shown over 10 iterations for the Experiments nos. 1 and 2. It can be easily seen that the overall character of these models does not really change much during the execution of the procedure. Not only this, they remain within a very clearly defined region of the two-dimensional parameter space. This can be, naturally, explained by the character of the data set (orientation of the data cloud), but also demonstrates the relative inertia of the calculations.

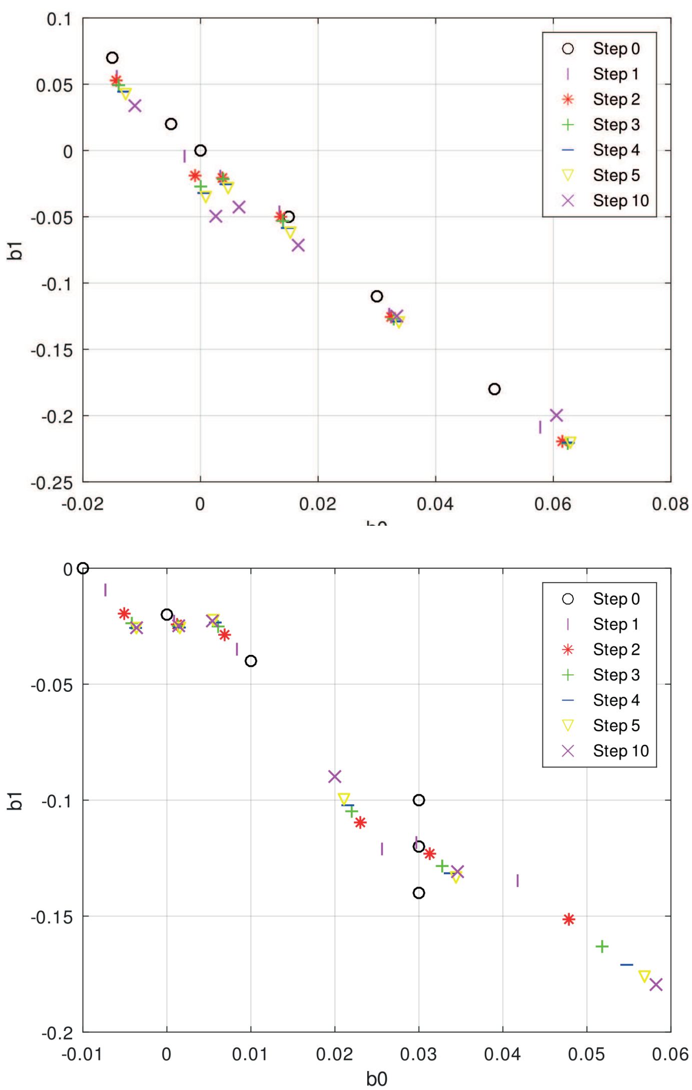

Figure 12. The changes of values of model coefficients over 10 iterations of the k-means-like procedure for Experiments nos. 1 and 2

Second, if we look at the correlation coefficient values, they appear to be somewhat haphazard, even if still much better than for the global model. This is, of course, at least partly due to the fact that the model sets are so different, and the municipality clusters, corresponding to models, are very diversified. Yet, if this is so, why the procedure does not lead to a possibly effective – even if not total – equalization of these values? Again: is this due to the character of this procedure, or rather results from the nature of data?

It appears that the procedure is struggling with the relatively flat landscape of the implicit objective function (total sum of errors), given that the data points form a homogeneous cloud, featuring definite shape and a relatively small subpopulation of evident outliers. Thus, one can consider quite a wide set of (potential) solutions as "satisfactory" on the basis, say, of the correlation coefficient values, the ultimate choice depending primarily upon the intuitive interpretation facility (see, in particular, the cases of Figs. 2, 3 and 11).

8. Conclusions

In this paper we describe an attempt of identifying separate models of migration for sub-populations of municipalities. The assumed form of the model of net migration is the linear function of unemployment level. Identification of the separate models proceeds through the procedure, which mimics the well-known k-means algorithm of clustering. The starting point for this procedure was first constituted by the classification of municipalities according to the magnitude and sign of error with respect to the single model for all municipalities. Given the dependence of the results from this procedure upon the starting point, a number of experiments were performed with different starting points (sets of models, corresponding to potential subgroups of municipalities).

The results obtained, even if showing an obvious improvement, especially over the single model for the whole population of municipalities, are somewhat disappointing in the perspective of expected convergence to an evident optimal solution. This effect may be due to two factors: the character of the data set (relatively homogeneous distribution of data points within the data cloud) and the specific features of the procedure (not ensuring convergence). Thus, we do not deal here with an "objective" optimal solution, but the results obtained may be treated as satisfying from the point of view of identification of subgroups of municipalities, for which some feasible interpretations, related to the respective models, can be forwarded, and even more – they may constitute the basis for corresponding policies.

This is very much like considering the task of clustering not as a task of "uncovering the existing objective inner structure of the data set", but rather as "dividing the data set into plausible subsets, which facilitate definite operations". Such "operations" might be purely intellectual (interpretation, understanding), but also practical – devising diversified policies. The task, therefore, becomes more of an operational research problem (which of the multiple obtained solution proposals satisfies the requirements, resulting from certain practical prerequisites?).

Bibliography

Basile, R., Girardi, A. and Mantuano, M. (2012) Migration and regional unemployment in Italy. *The Open Urban Studies Journal*, 5, 1-13. http://dx.doi.org/10.2174/1874942901205010001.

Basile, R., Mantuano, M., Girardi, A. and Russo, G. (2019) Interregional migration of human capital and unemployment dynamics: Evidence from Italian provinces. *German Economic Review*, **20**(4), 1-30. https://doi.org/10.1111/geer.12172.

Baumann, R., Svec, J. and Sanzari, F. (2015) The relationship between net migration and unemployment: The role of expectations. *Eastern Economic Journal*, 41, 443-458. https://doi.org/10.1057/eej.2014.44.

Blanchflower, D. G. and Shadforth, C. (2009) Fear, unemployment and migration. *Economic Journal*, **119**(535), F136-F182.

Card, D. (2001) Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, **19**(1), 22-63. https://doi.org/10.1086/209979.

Docquier, F., Peri, G. and Ruyssen, I. (2014) The Cross-country Determinants of Potential and Actual Migration. *International Migration Review*, **48**(s1), S37-S99. https://doi.org/10.1111/imre.12137.

Dustman, C., Fabbri, F., Preston, I. and Wadsworth, J. (2003) *The local labour market effects of immigration in the UK*, Home Office Online Report 06/03, London: Research Development and Statistics Directorate, Home Office.

Espinosa, A. M. and Díaz-Emparanza, I. (2021) The Long-term Relationship Between International Labour Migration and Unemployment in Spain. *Journal of International Migration and Integration*, 22, 145-166. https://doi.org/10.1007/s12134-019-00716-6.

Fendel, T. (2014) Work-related Migration and Unemployment. *Journal for Labour Market Research*, 47, 233-243.

Granato, N., Haas, A., Hamann, S. and Niebuhr, A. (2015) The Impact of Skill Specific Migration on Regional Unemployment Disparities in Germany. *Journal of Regional Science*, 55, 513-539. https://doi.org/10.1111/jors.12178.

Günther, J. (2000) Regional unemployment differentials and internal migration in Hungary. *SEER - South-East Europe Review for Labour and Social Affairs*, **3**(1), 135-145.

Heid, B. and Larch, M. (2012) Migration, Trade and Unemployment. *Economics: The Open-Access, Open-Assessment E-Journal*, 6 2012-4, http://dx.doi.org/10.5018/economics-ejournal.ja.2012-4.

Herzog, H. W., Schlottmann, A. M. and Boehm, T. P. (1993) Migration as Spatial Job–search: A Survey of Empirical Findings. *Regional Studies,* **27**(4), 327–340. https://doi.org/10.1080/00343409312331347595.

Huynh, H. H. and Vo, D. H. (2023) The Effects of Migration on Unemployment: New Evidence from the Asian Countries. *Sustainability*, 15, 11385. https://doi.org/10.3390/su151411385.

Jennissen, R. (2003) Economic Determinants of Net International Migration in Western Europe. *European Journal of Population*, 19, 171-198.

Jofre-Monseny, J. (2014) The effects of unemployment protection on migration in lagging regions. *Journal of Urban Economics*, 83, 73-86. http://dx.doi.org/10.1016/j.jue.2014.08.001.

Kanga, D., Loko, B., Agou, G. and Kpodar, K. (2024) *Intra-African Migration: Exploring the Role of Human Development, Institutions, and Climate Shocks*, WP/24/97, International Monetary Fund.

Kilic, C., Yucesan, M. and Ozekicioglu, H. (2019) Relationship Between Migration and Unemployment: Panel Data Analysis for Selected OECD Countries. *Montenegrin Journal of Economics*, **15**(3), 101-111. DOI: 10.14254/1800-5845/2019.15-3.7.

Kondoh, K. and Yabuuchi, S. (2012) Unemployment, environmental policy, and international migration. *The Journal of International Trade & Economic Development*, **21**(5), 677-690. https://doi.org/10.1080/09638199.2010.535613.

Kulkolkarn, K. and Potipiti, T. (2007) Migration, Wages and Unemployment in Thailand. *Chulalongkorn Journal of Economics*, **19**(1), 1-22. https://doi.org/10.1007/978-3-030-31527-6.

Lee, E. S. (1966) A Theory of Migration. *Demography.* 3, 47-57.

Lemos, S. and Portes, J. (2008) New Labour? The Impact of Migration from Central and Eastern European Countries on the UK Labour Market. *IZA Discussion Paper* No. 3756. http://dx.doi.org/10.2139/ssrn.1286694.

MacQueen, J. B. (1967) Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1. University of California Press, 281–297.

Mbachu, H. I. (2022) Modelling the Effect of Gross Domestic Product (GDP) Per Capita on Population Growth, Youth Unemployment and Net Migration. *International Journal of Mathematics and Statistics Invention*, **10**(1), 12-21.

MENG, X. AND ZHANG, D. (2010) *Labour Market Impact of Large Scale Internal Migration on Chinese Urban 'Native' Workers,* Discussion Paper No. 5288, Bonn: The Institute for the Study of Labor. https://docs.iza.org/dp5288.pdf (accessed 12.06.2024).

OWSIŃSKI J. W., STAŃCZAK J., ŚLESZYŃSKI P. AND WIŚNIEWSKI R. (2024) Modeling Socio-Economic Phenomena: Some Critical Thoughts and a Positive Example. In: C. Ciurea, P. Pocatilu, F. G. Filip, eds., *Proceedings of 23rd International Conference on Informatics in Economy (IE 2024). Recent Results in Education, Research and Applications. Actions. Smart Innovations, Systems and Technologies* **SIST 426**. Springer Nature Singapore Pte Ltd.

PIORE, M. J. (1979) *Birds of Passage: Migrant Labour and Industrial Societies.* Cambridge: Cambridge University Press.

ŚLESZYŃSKI, P. (2016). A synthetic index of the spatio-temporal accessibility of communes in Poland. *Geographia Polonica*, **89**, 4, 567-574. http://dx.doi.org/10.7163/Gpol.0074

ŚLESZYŃSKI, P. (2024) Migracje wewnętrzne [Internal migrations]. In: Bojarska-Lis (ed.), *Sytuacja demograficzna Polski, Raport 2022-2023*, 113-134, Warszawa: Rządowa Rada Ludnościowa.

STAŃCZAK, J., OWSIŃSKI, J., ŚLESZYŃSKI, P. AND WIŚNIEWSKI, R. (2024) Modeling internal migrations in Poland at municipality level [dataset], Mendeley Data, v2, https://data.mendeley.com/datasets/wyjzkfx6yd/3.

STEINHAUS, H. (1957) Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.* (in French). **4** (12): 801–804. MR 0090073. Zbl 0079.16403.

WHITE, A. (2015) Informal Practices, Unemployment, and Migration in Small-Town Poland. *East European Politics and Societies*, **30**(2), 404-422.

ZABEL, J. E. (2012) Migration, housing market, and labor market responses to employment shocks. *Journal of Urban Economics*, **72**(2-3), 267-284. https://doi.org/10.1016/j.jue.2012.05.006