Исходный набор данных ***D*** состоит из некоторого числа муниципальных образований , каждая из которых представлена данными за определенный промежуток времени . Таким образом, следует выразить как кортеж следующего вида:

,

где – является социально-экономическим состоянием в момент времени . Все состояния являются неотъемлемой частью набора данных

В свою очередь есть упорядоченный набор социально-экономических индикаторов, также представляющим собой кортеж вида:

,

где – это определенный социально-экономический индикатор.

Каждое состояние сопровождается некоторым миграционным сальдо ***s***, которое вычисляется как:

,

где **inflow** количество прибывших людей, а **outflow** – количество выбывших.

При этом делается обоснованное допущение, что значения **inflow** и **outflow** имеют зависимость от , что можно представить некоторой функцией миграционной привлекательности :

.

Соответственно, задача исследования заключается в разработке метода определения необходимого вектора изменений для потенциального обеспечения миграционной привлекательности, то есть:

В первую очередь необходимо разделить элементы датасета ***D*** на некоторое количество подмножеств , именуемых кластерами, по принципу их похожести. То есть на основании заданного алгоритма формируются кластеры

где – индекс кластера, – общее число кластеров, а  **–** индекс принадлежности некоторого к .

Далее осуществляется разбивка всех сформированных кластеров на положительный и отрицательный субкластеры , каждая из которых включает в себя состояния , удовлетворяющие два определенных условия:

*.*

Из выражения видно, что положительный субкластер кластера включает в себя только те состояния , которые продемонстрировали положительное сальдо. В свою очередь будет отличаться только условием . Совершенно очевидно, что .

В качестве направления вектора развития можно использовать медианные значения факторов положительного субкластера, описанного в виде кортежа:

,

где – медианное значение фактора . Вычисление разницы является прямым способом оценки состояния с точки зрения пространства миграционной привлекательности. При этом возникает возможность построить вектора развития на основе для любого

,

где – это измененный фактор по сравнению с исходным согласно следующим условиям:

*,*

при этом из выражения понятно, что не будет отличаться от исходного значения , если этот фактор больше медианы . Несмотря на то, что замена в меньшую сторону в данном случае может рассматриваться как сокращение доходов, а значит нечто позитивное, на данном этапе исследования вектор развития выстраивается по принципу «не навреди».

*……..*

***Кластеризация.***

***…***

После кластеризации набора данных ***D*** на ***n*** кластеров, возникает возможность оценить как медианное значение миграционного сальдо ***s***,так и долю примеров с отрицательным и положительным сальдо в каждом кластере.

На рисунке 1a на боксплот графике видно, что нет ни одного кластера, имеющего положительную медиану (разделительная линия на блоке) для миграционного сальдо, что, несомненно, говорит о превалировании тенденции оттока населения при большинстве состояний муниципальных образований [???]. Это также ярко демонстрируется на столбчатом графике, котором показаны доли примеров с отрицательным миграционным сальдо в кластере.



Рис. 1. Net migration analysis in each cluster

Однако, рассмотренный столбчатый график (рис. 1б) в том числе демонстрирует наличие положительных состояний с точки зрения миграционного сальдо ***s***. Так, в лучшем случае доля примеров с положительным сальдо составила ~35% в кластере 4, а в худшем ~15% в кластере 5 (см. рис. 1б). Таким образом, можно сделать вывод, что, во-первых, миграционно привлекательные муниципальные образования или как минимум только определенные их состояния существуют в каждом кластере, а, во-вторых, это даёт возможность рассмотреть их с позиции разницы социально-экономических характеристик их состояний .

В свою очередь анализ этого различия является одним из фундаментальных составляющих для достижения поставленной цели, а именно разработки метода определения вектора развития для обеспечения миграционной привлекательности. Так, этот анализ позволит:

- оценить значимость отдельных элементов (факторов) состояний на положительное миграционное сальдо как в целом, так и внутри конкретных кластеров;

- оценить наличие устойчивых поведенческих зависимостей элементов (факторов) состояний с точки зрения природы их воздействия.

***Анализ разницы между положительными и отрицательными субкластерами.*** Для реализации вышеупомянутого анализа необходимо разделить каждый кластер на два субкластера, а именно «положительный» и «отрицательный». В положительный субкластер входят только те состояния, при которых было зафиксировано миграционное сальдо больше 0, а в отрицательный наоборот.

Это даёт нам возможность вычислить медианы каждого фактора для этих субкластеров. После этого можно рассчитать насколько фактор в положительном кластере отличается от фактора в отрицательном.

Важно отметить, что миграционное сальдо и численность населения не использовались в качестве факторов при кластеризации, то есть кластеры были сформированы алгоритмом к-средних независимо от этих характеристик.

Итак, на диаграмме ниже показаны различия медиан положительного субкластера от отрицательного по всем кластерам для каждого элемента (фактора) состояния . Если значение меньше 0%, то это говорит о том, что этот фактор в отрицательном субкластере больше, нежели в положительном.

На графике отчётливо наблюдаются некоторые зоны синхронного повышения значений и их уменьшения. Эти зоны демонстрируют наиболее значимые факторы, поскольку во всех кластерах их большее значение ведет к положительному сальдо. К таким признакам относятся retailturnover (сред. разн. 11%), harvest (сред. разн. 17%) и factoriescap (сред. разн. 18%). Почти такое же влияния имеет фактор agrprod (сред. разн. 15%), который только в кластере №3 имеет больше значение в отрицательном субкластере (см. рис. 2, фиолетовая линия). Похожая ситуация наблюдается и с avgsalary (сред. заработная плата), однако, разница небольшая и составляет приблизительно 4%.



Также серьезный интерес вызывает разница между положительным и отрицательным субкластером по факторам «beforeschool» и «hospitals». На графике видно, что в 5, 4 и 0 кластере данные факторы в среднем имеют большее значение в отрицательном субкластере, то есть наблюдается большая миграция при низких значениях этих факторов.

На первый взгляд этот результат может показаться противоречивым. Однако, одним из главных объяснений может являться разный тип миграции [111], который наблюдается в кластерах. Так, в РФ сильно превалирует молодежная миграция (когорта 15-19), которая переезжает в целях учёбы [111]. Несомненно, молодое поколение также мигрирует и в расчете карьерных перспектив. Таким образом, для молодых возрастных групп, не имеющих детей, здравоохранение и дошкольные организации не являются важными факторами при принятии решения о переезде. Однако, в эти кластеры (5, 4 и 0) характеризуются высокими значениями foodseats, retaliturnover, harvest, agrprod, livestock что также согласуется с интересами молодежной и рабочей миграции.

При этом в РФ ярко наблюдается тенденция миграции людей с детьми дошкольного возраста, а также так называемая пенсионная миграция [111]. Обратите внимание на красную, зеленую и фиолетовую линию графика (см. рис. 2, кластеры 1, 2 и 3), которые вероятнее всего отражают эту тенденцию, то есть в этих случаях количество мест в дошкольных организациях и количество лечебных организаций играет существенную роль при миграции. Также эти 3 кластера характеризуются положительной миграцией в факторе спортивных сооружений (sportsvenue), что даёт возможность устроить детей на какие-либо спортивные секции.

Таким образом, полученная разница по отдельным факторам вполне может объясняться различным типом миграции, которая наблюдается в стране [111]. Однако, нельзя также исключать, например, причину с точки зрения качества, а не количества.

Следует еще отметить, что хоть численность населения и не учувствовала в процессе кластеризации, при её рассмотрении она тоже является сильным признаком, то есть медианное значение в положительном субкластере во всех случаях больше, чем в отрицательном (см. рис. 2). Этот результат согласуется с гравитационной территорией миграционных потоков [???].

[111] Karachurina, L.B., Mkrtchyan, N.V. To the City or to the Suburbs: What Russians Choose at Different Stages of Life Course. Reg. Res. Russ. 14 (Suppl 1), S55–S66 (2024). https://doi.org/10.1134/S2079970524600689

Помимо кластерного анализа, позволившего идентифицировать как разницу социально-экономического состояния в отдельных суб-кластерах, так и степень их вероятного влияния на миграционные потоки, что позволяет сформировать некоторый вектор развития территории, возникает возможность получения косвенной оценки, подтверждения этого вектора на основе модели прогнозирования миграции. Подобная модель была разработана авторами исследования и представлена в работе **[наша].**

Сущность модели заключается в том, что зная социально-экономического состояние в момент времени , а также миграционное сальдо в момент времени , можно создать простую модель обучения с учителем для прогнозирования миграционного сальдо в момент (следующий год) при определенном состоянии .

Соответственно, для реализации данной оценки необходимо обучить модель на тех же признаках, что и использовались при кластерном анализе. Однако, поскольку прогнозируется следующий год, то количество примеров для обучения сокращается в виду того, что, во-первых, исключаются все примеры 2022 года, поскольку нет данных за 2023, а также у отдельных есть пропуски, то есть нет данных за какой-то определенный год. Итоговый датасет для обучения модели прогноза включает в себя чуть больше 7000 примеров.

Для проведения экспериментально оценки было отобрано по 5 различных примеров из каждого кластера, демонстрирующих миграционный отток на следующий год. Эти примеры будут изменены согласно ранее описанному алгоритму. В таблице ниже показаны некоторые примеры. В синих строках указаны реальные значения, а в зеленных измененные. Если по какому-то признаку отсутствует изменение (см. табл. 1, пример Pavlovsky, признак f8), то это говорит, что в данном примере это значение лучше, чем медиана положительного субкластера.

Таблица 1. Отобранные примеры для исследования

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **oktmo** | **year** | **name** | **clust** | **…** | **f5** | **…** | **f8** | **f9** | **…** | **f15** | **net mig (next year)** |
| **actual** | 22719000 | 2021 | Vorotynsky | 0 | … | 729883 | … | 8 | 39 | … | 236233 | -140 |
| **changed** | 22719000 | 2021 | Vorotynsky | 0 |  | 1094086 |  | 78 | 39 |  | 5013077 | … |
| **actual** | 3639000 | 2019 | Pavlovsky | 0 | … | 2824761 | … | 92 | 25 | … | 12405299 | -476 |
| **changed** | 3639000 | 2019 | Pavlovsky | 0 | … | 1094086 | … | 92 | 30 | … | 12405299 | … |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| **actual** | 1625000 | 2021 | Loktevsky | 5 | … | 573514 | … | 60 | 28 | … | 828015 | -294 |
| **changed** | 1625000 | 2021 | Loktevsky | 5 | … | 786293 | … | 71 | 28 | … | 3002048 | … |
| **actual** | 60730000 | 2021 | Novoshakhtinsk | 5 | … | 3389505 | … | 249 | 22 | … | 5939939 | -429 |
| **changed** | 60730000 | 2021 | Novoshakhtinsk | 5 | … | 3389505 | … | 249 | 26 | … | 5939939 | … |

The dataset for the research was collected from open source data provided by Rosstat. It contains information on ~1800 municipalities for the period of 2014 to 2022. In the majority of cases (approx. ~80%) the municipality is a district. Despite the fact that there are more than 1000 cities in Russia, the federal agency mainly does not provide data on small or rural settlements separately in open access.

The number of socio-economic indicators is 16. These were chosen as suspicious of influencing migration [..., наша статья]. The list of features f is as follows:

1. Number of people - popsize (ppl.)

2. Average number of employers in organizations - avgemployers (ppl.)

3. Average salary - avgsalary (rub.)

4. Shopping area - shoparea (sq.m.)

5. Number of seats in café, bars and restaurants - foodseats (num.)

6. Retail turnover - retailturnover (thnd. rub.)

7. Live area per capita - livarea (sq.m.)

8. Number of sporting venues - sportsvenue (num.)

9. Number of services (barbershops, repairs) - servicesnum (num.)

10. Length of roads - roadslen (km.)

11. Number of livestock - livestock (num.)

12. Productivity of land (vegetables) - harvest (centners)

13. Agricultural production - agrprod (thnd. rub.)

14. Number of healthcare organizations - hospitals (num.)

15. Number of places in preschool organizations - preschool (num.)

16. Volume of self-produced goods - factoriescap (thnd. rub.)

Thus, the general dataset ***D*** contains a certain number of municipalities ***M***, with each of them represented by data for a defined period ***t*** (in years). The ***M*** may be described as a tuple:

,

where – is the socio-economic state of during time period . All states of are examples of dataset

Additionally, is an ordered set of socio-economic indicators that define a particular state of ***M,*** and also can be represented as a tuple:

,

where – is a specific socio-economic indicator.

Each state is characterised by a certain net migration ***s***, which could be calculated as follows:

,

where **inflow** – number of people entering the area, **outflow** – number of people leaving the area.

It is possible to make fair assumption [???] that the number of ***inflows*** and ***outflows*** depends on , which can be described by a certain function of migration attractiveness ***mig***:

.

Therefore, the main goal of the research is to develop a method to define the vector of necessary changes for providing the attractiveness of migration:

The first step is to divide the elements of the dataset ***D*** onto a specific number of subsets (clusters) according to their similarity. Thus, based on a certain algorithm, clusters will be formed:

&

where – index of the cluster, – number of clusters, а  **–** index of affiliation for specific to .

The next step is to separate all produced clusters on positive and negative sub-clusters , each of which includes particular states that are satisfying two conditions:

*.*

From the conditions presented, it can be seen that the positive sub-cluster of cluster contains only those states of the municipality that had a positive net migration. On the other hand, is defined by the inverse condition . It is also obvious that .

The median values of the factors from the positive sub-cluster can be used as a relative vector of the development for similar municipalities. These set of medians could described as a tuple:

,

where  **–** the median value of the factor . It is possible to use the difference of as a straightforward approach to evaluate the state according to the space of migration attractiveness. Thus, it will allow to create the development vector for any based on :

,

where – the modified factor compared to the original according to following conditions:

*,*

from the above conditions, it is clear that will not differ from its initial value if this factor is indeed greater than the median . Despite the fact that in this case replacement to the smaller side could theoretically be beneficial (because of cost reduction), at this stage of the research the development vector will be built on the principle of ‘do no harm’.

It is important to note that this approach has an obvious flaw, which occurs when a specific feature is significantly smaller than the median of the positive sub-cluster. In this case, it will be unrealistic to recommend such a development vector. The proposed solution is to use the percentage difference between the median value of the feature in the positive and negative sub-clusters as the recommended growth:

Across all clusters (see Figure 3), this difference does not exceed 61%. Figure 3 also shows negative percentage differences, which basically indicates that . In this particular context, the principle of “do no harm” is also respected:

The threshold value for is set to 100%. In other words, no modifications will be more than 100% for the recommended development vector:

***Внутрикластерный вектор развития.***

After clustering the dataset ***D*** into ***6*** clusters, it is necessary to first evaluate the relevance of the resulting partition. The most straightforward approach is to visualize the differences between clusters by comparing the median values of each feature. The radar diagram demonstrates that the clusters differ significantly from one another (Fig. 1).



Fig. 1. Median values of each feature in different clusters (each value is normalized from 0 to 1)

It is clear that Cluster 1 (see Fig. 1, red area) shows the highest values of “shoparea”, “foodseats”, “retailturnover”, and “servicesnum” among all clusters and probably includes the most developed municipalities in terms of resident comfort. On the other hand, Cluster 4 (see Fig. 1, bright blue area) shows high levels of “avgsalary”, “avgemployers”, “factoriescap”, and “harvest”, which likely represent municipalities with developed industrial sectors.

Therefore, it can be concluded that the clustering process is adequate, as the resulting clusters demonstrate clear distinctions.

…

Following the clustering process of the dataset ***D*** into ***n*** clusters, it is possible to analyse both the medians of net migration ***s*** and the proportion of negative examples (***s*** < 0) within each cluster.

As illustrated by the boxplot of net migration (Fig. 1a), it is evident that the medians (the lines dividing the boxes) of all clusters are negative. This indicates the prevalence of population outflow in the majority of states for settlements ***M***. This is also confirmed by the bar chart in Figure 1b, which shows the proportion of negative examples (***s <*** 0) for all clusters.

However, the bar chart (Fig. 1) also demonstrates positive examples in terms of net migration. In the best case the proportion of positive examples is approximately ~35% in cluster 4, while in the worst case it is around ~15% in cluster 5 (Fig.1 b).

The following conclusions can be made: firstly, within each cluster there are migration-attractive settlements ***M***, or at the very least, their particular states ; and secondly, it is possible to assess the difference between positive and negative states in terms of socio-economic factors within each cluster.

The analysis of these differences is a fundamental step in accomplishing the main goal of the research, which is to develop a method for identifying a vector of changes that has the potential to increase the migration attractiveness of specific settlements. On the basis of such an analysis, the following aspects can be assessed:

* the importance of certain elements (factors) of the states on positive net migration, both in general and within specific clusters;
* the presence of stable behavioral dependencies of certain elements (factors) of the states in terms of the nature of their impact.

***The difference between positive and negative sub-clusters.*** In order to implement the analysis that mentioned above, it is necessary to divide each cluster into two sub-clusters: “positive” and “negative”. The positive sub-cluster includes only those states m\_t that are characterised by net migration s > 0, while the negative sub-cluster is characterised by the opposite condition.

The next step is to calculate the median of each element (factor) of the states m\_t for specific sub-clusters. Based on these medians, an assessment can be made of how much each factor in the positive sub-cluster differs from the negative one.

It should be noted that neither net migration nor population size (popsize) is used as a factor in the process of clusterization. Consequently, the formation of clusters is entirely independent of these characteristics.

The diagram below (Fig. 2) demonstrates the difference between the medians of socio-economic factors (elements of ***m\_t***) from the positive and negative sub-clusters for each cluster. The value less than 0% indicates that this factor is greater in the negative sub-cluster compared to the positive sub-cluster.

Firstly, the diagram shows some areas of overall increase and decrease in the difference, illustrating the factors that have a similar nature of impact in all clusters (Fig. 2). The factors that demonstrate a common increase in all cases should be considered the most significant, as their higher values consistently result in positive net migration. These include retailturnover (avg. diff. ~11%), harvest (avg. diff. ~17%) and factoriescap (avg. diff. ~18%). The factor agrprod (avg. diff. ~15%) has almost the same influence, except cluster 3 (see Fig. 2, purple line). A similar situation is observed with avgsalary, however, the average difference is quite small ~ 4%.

It is important to note that although the population size was not involved in the clustering process, its higher value always led to positive net migration too (see Fig. 1). This result is actually consistent with the gravitational theory of migration flows [???].

The difference between the positive and negative sub-clusters by the factors ‘beforeschool’ and ‘hospitals’ should be considered carefully. The diagram shows that in 5, 4 and 0 clusters these factors actually have a higher value in the negative sub-cluster. This indicates that there is a positive net migration when these factors are lower.

At first glance, this result may seem contradictory. However, one of the main explanations may be the different type of migration [111] observed in the clusters. Thus, youth migration (cohort 15-19), which moves for educational purposes, strongly prevails in the Russia Federation [111]. Undoubtedly, the younger generation also migrates for in the reason of career opportunities. So, for younger age groups without children, healthcare and per-school organizations are not important factors in decision to move. Also these clusters (5, 4 and 0) are characterized by high values of foodseats, retaliturnover, harvest, agrprod, livestock, which is consistent with the interests of youth and labour migration.

At the same time, there is also a clear trend of family migration with pre-school age children and so-called retirement migration [111]. The red, green and purple lines of the diagram (see Fig. 1, clusters 1-3), illustrates this trends most clearly around "hospitals" and "beforeschool" factors. In such cases, healthcare and per-school organizations are significant factors in decision-making. In addition, these three clusters are distinguished by positive migration when “sportsvenue” is higher, indicating the availability of opportunities for children to engage in sports.

However, it is obvious that not only the types of migration in the country [111] could explain such a difference. For instance, the nature of the factors is quantitative, while the qualitative aspects are also significant (fewer hospitals, but better in terms of medical equipment etc.).

…

The simplified scheme presented in Figure 4 describes the proposed approach.

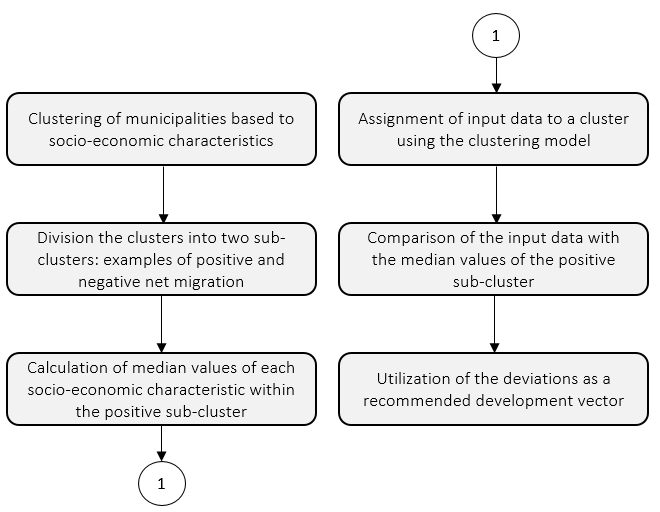


Fig. 4. The main steps of the proposed approach

In addition to the cluster analysis that made it possible to define a specific development vector for the area by identifying both the differences in socio-economic state in each cluster and their likely impact on migration flows, it is possible to make an indirect confirmation of this vector using a migration forecasting model. Such a model has been developed by the authors of the research and published in […]. The model can be summarized as follows: given the socio-economic state at time and the net migration at time (i.e. the following year),a simple supervised learning model can be built to predict the net migration at time given any state .

In order to implement such an assessment of development vector, it is necessary to train the model with the same features used in the cluster analysis. It should be noted that because of the forecasting process for the subsequent year, there will be a significant reduction in the number of training examples. This is primarily due to the exclusion of all examples from 2022, as no data on net migration is available for 2023, and the lack of data for particular years. The overall dataset consists of approximately 7000 examples.

A random forest-based model is proposed for migration forecasting. In addition to being simple and flexible, this approach has already been tested in previous studies on migration forecasting [наша статья]. Random forest is also a valuable tool for evaluating feature importance (according to mean decrease in impurity), which is particularly relevant as it allows for comparison with the results of cluster analysis.

It is not anticipated that the feature importance will be identical, as the forecasting model predicts the number of migrants, while the cluster analysis focused on the difference between positive and negative net migration. Nevertheless, some similarities may reasonably be expected.

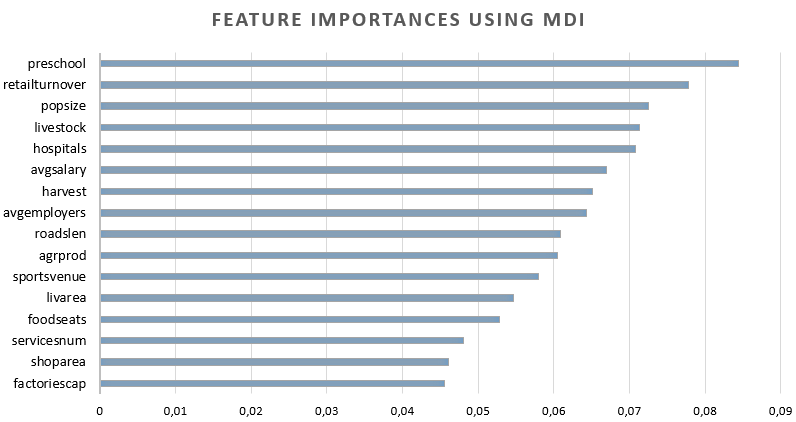


Fig. 5. Random forest feature importances according to mean decrease in impurity

Five different examples characterized by a migration outflow in the following year were selected from each cluster for experimental evaluation of the development vector. These examples will be modified according to the method described earlier.

Table 1 presents some of these examples. The blue rows represent the actual values, while the green rows show the modified values. The absence of a change for a given feature (see Table 1, Pavlovsky, f8) indicates that in these examples the value is already greater than the median of the positive sub-cluster.

Таблица 1. Отобранные примеры для исследования

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Actual** | **Modified** | **.** | **Actual** | **Modified** |
| **oktmo** | 22719000 | 22719000 | **.** | 60730000 | 60730000 |
| **year** | 2021 | 2021 | **.** | 2021 | 2021 |
| **name** | Vorotynsky m.d. | Vorotynsky m.d. | **.** | Novoshakhtinsk | Novoshakhtinsk |
| **clust** | 0 | 0 | **.** | 5 | 5 |
| **f1** | 2414 | 3074,9 | **.** | 11442 | 11442 |
| **f2** | 19887,4 | 20815,4 | **.** | 18218,3 | 20804,7 |
| **f3** | 8105 | 9990,9 | **.** | 69421,6 | 69421,6 |
| **f4** | 474 | 881 | **.** | 80 | 104,05 |
| **f5** | 729883,9 | 1094086,8 | **.** | 3389505 | 3389505 |
| **f6** | 39,4 | 39,4 | **.** | 21,8 | 26,1 |
| **f7** | 45 | 52,2 | **.** | 165 | 165 |
| **f8** | 8 | 9,8 | **.** | 249 | 249 |
| **f9** | 296,9 | 338,9 | **.** | 369,1 | 497,8 |
| **f10** | 32144 | 35709 | **.** | 18257 | 19594,9 |
| **f11** | 60102,4 | 73632,7 | **.** | 57790 | 57790 |
| **f12** | 550648,1 | 744853,9 | **.** | 521260,1 | 708166,2 |
| **f13** | 23 | 35,7 | **.** | 6 | 6 |
| **f14** | 946 | 1599,1 | **.** | 3403 | 3403 |
| **f15** | 236233,5 | 356159,9 | **.** | 5939939,1 | 5939939,1 |
| **net migration (2022)** | -140 | -94 (predicted) | **.** | -429 | -315 (predicted) |

There is no doubt that it is necessary to forecast not only the modified examples, but also the actual ones in order to evaluate the behavior of the model. In the diagram below (Fig. 3), the blue line represents the actual net migration, and the black line shows the model’s prediction based on the original features. It is clear that model is not perfect and mainly biased toward the positive side.

Under such conditions, it will be appropriate to compare the predictions of the modified examples with the predications for unmodified ones in order to assess the impact of the development vector on migration attractiveness. The green line shows the predicted net migration for the modified examples (Fig. 3).

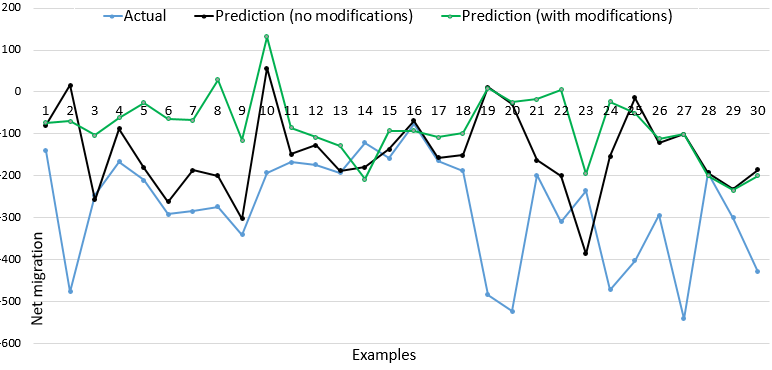


Fig 3. Actual net migration compared to predicted one

The diagram illustrates that in 70% of cases (21 out of 30), modified examples show less migration outflow. The average predicted net migration for the unmodified examples is -147, while for the modified ones it is -83 (1.7 times less).

These results indirectly confirms that the development vector, as defined by the clustering analysis, may indeed contribute to positive changes in migration flows.