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:

*……..*

***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%.



Интересно 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 [???].

***Внутрикластерный вектор развития.***

…

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%.