КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 6 С. 1501–1513
DOI: 10.20537/2076-7633-2020-12-6-1501-1513

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.237.8

Оценка качества кластеризации панельных данных с использованием методов Монте-Карло (на примере данных российской региональной экономики)

И. Л. Кирилюк[1,a], О. В. Сенько[2]

[1] Институт экономики Российской академии наук,
Россия, 117218, г. Москва, Нахимовский проспект, д. 32

[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук,
Россия, 119333, г. Москва, ул. Вавилова, д. 44/2

E-mail: [a] igokir@rambler.ru

Получено 04.05.2020, после доработки — 02.09.2020.
Принято к публикации 18.09.2020.

В работе рассматривается метод исследования панельных данных, основанный на использовании агломеративной иерархической кластеризации — группировки объектов на основании сходства и различия их признаков в иерархию вложенных друг в друга кластеров. Применялись 2 альтернативных способа вычисления евклидовых расстояний между объектами — расстояния между усредненными по интервалу наблюдений значениями и расстояния с использованием данных за все рассматриваемые годы. Сравнивались 3 альтернативных метода вычисления расстояний между кластерами. В первом случае таким расстоянием считается расстояние между ближайшими элементами из двух кластеров, во втором — среднее по парам элементов, в третьем — расстояние между наиболее удаленными элементами. Исследована эффективность использования двух индексов качества кластеризации — индекса Данна и Силуэта для выбора оптимального числа кластеров и оценки статистической значимости полученных решений. Способ оценивания статистической достоверности кластерной структуры заключался в сравнении качества кластеризации, на реальной выборке с качеством кластеризаций на искусственно сгенерированных выборках панельных данных с теми же самыми числом объектов, признаков и длиной рядов. Генерация производилась из фиксированного вероятностного распределения. Использовались способы симуляции, имитирующие гауссов белый шум и случайное блуждание. Расчеты с индексом Силуэт показали, что случайное блуждание характеризуется не только ложной регрессией, но и ложной кластеризацией. Кластеризация принималась достоверной для данного числа выделенных кластеров, если значение индекса на реальной выборке оказывалось больше значения 95%-ного квантиля для искусственных данных. В качестве выборки реальных данных использован набор временных рядов показателей, характеризующих производство в российских регионах. Для этих данных только Силуэт показывает достоверную кластеризацию на уровне $p < 0.05$. Расчеты также показали, что значения индексов для реальных данных в целом ближе к значениям для случайных блужданий, чем для белого шума, но имеют значимые отличия и от тех, и от других. Визуально можно выделить скопления близко расположенных друг от друга в трехмерном признаковом пространстве точек, выделяемые также в качестве кластеров применяемым алгоритмом иерархической кластеризации.

Ключевые слова: достоверность кластеризации, панельные данные, мезоэкономика, экономика регионов

COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 6 P. 1501–1513
DOI: 10.20537/2076-7633-2020-12-6-1501-1513

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.237.8

Assessing the validity of clustering of panel data by Monte Carlo methods (using as example the data of the Russian regional economy)

I. L. Kirilyuk[1,a], O. V. Senko[2]

[1] Institute of Economics, Russian Academy of Sciences,
32 Nakhimovskii pr., Moscow, 117218, Russia

[2] Federal Research Center Computer Science and Control, Russian Academy of Sciences,
44/2 Vavilova st., Moscow, 119333, Russia

E-mail: [a] igokir@rambler.ru

Received 04.05.2020, after completion — 02.09.2020.
Accepted for publication 18.09.2020.

The paper considers a method for studying panel data based on the use of agglomerative hierarchical clustering — grouping objects based on the similarities and differences in their features into a hierarchy of clusters nested into each other. We used 2 alternative methods for calculating Euclidean distances between objects — the distance between the values averaged over observation interval, and the distance using data for all considered years. Three alternative methods for calculating the distances between clusters were compared. In the first case, the distance between the nearest elements from two clusters is considered to be distance between these clusters, in the second — the average over pairs of elements, in the third — the distance between the most distant elements. The efficiency of using two clustering quality indices, the Dunn and Silhouette index, was studied to select the optimal number of clusters and evaluate the statistical significance of the obtained solutions. The method of assessing statistical reliability of cluster structure consisted in comparing the quality of clustering on a real sample with the quality of clustering on artificially generated samples of panel data with the same number of objects, features and lengths of time series. Generation was made from a fixed probability distribution. At the same time, simulation methods imitating Gaussian white noise and random walk were used. Calculations with the Silhouette index showed that a random walk is characterized not only by spurious regression, but also by "spurious clustering". Clustering was considered reliable for a given number of selected clusters if the index value on the real sample turned out to be greater than the value of the 95% quantile for artificial data. A set of time series of indicators characterizing production in the regions of the Russian Federation was used as a sample of real data. For these data only Silhouette shows reliable clustering at the level $p < 0.05$. Calculations also showed that index values for real data are generally closer to values for random walks than for white noise, but it have significant differences from both. Since three-dimensional feature space is used, the quality of clustering was also evaluated visually. Visually, one can distinguish clusters of points located close to each other, also distinguished as clusters by the applied hierarchical clustering algorithm.

Keywords: clustering validity, panel data, mesoeconomics, regional economics

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 6, pp. 1501–1513 (Russian).

1. Введение

Во многих областях знания исследователи имеют дело с наборами временных рядов, или панельных данных, объединенных в подгруппы на основании каких-либо признаков. Иногда такое разбиение множества временных рядов на подгруппы не является изначально очевидным, и его выявление является предметом интереса исследователей. В этом случае возникает задача кластеризации временных рядов [Liao,2005; Aghabozorgi et al., 2015; Ивахненко и др., 2007]. В некоторых публикациях рассматривается также задача кластеризации панельных данных, например в [Kapetanios, 2006; Niu, 2012]. Традиционные подходы в анализе панельных данных во многом сосредоточены на проверке гипотез о равенстве коэффициентов в панелях, задача их кластеризации редко ставится, хотя она в каких-то случаях может дать более детальную, существенную информацию. А игнорирование кластерной структуры данных может повлечь некорректные статистические выводы при их исследовании.

Важным фактором в задачах кластеризации является оценка ее качества и достоверности. Для оценки качества кластеризации предложено множество разнообразных индексов (в литературе есть перечни из десятков таких индексов [Halkidi et al., 2001; Charrad et al., 2014; Сивоголовко, 2011]). Эти индексы позволяют сравнивать различные варианты кластеризации. С их помощью определяется, разбиение на какое число кластеров в исследуемом объеме признакового пространства дает наиболее достоверную и выраженную группировку. Однако они не позволяют напрямую сделать вывод о достоверности полученных решений. Для оценки достоверности кластеризации существуют различные подходы. Например, вывод о ее достоверности может быть сделан на основании мнения экспертов или при соответствии кластеризации значениям каких-то внешних факторов, не используемых при кластеризации. Одним из важных способов установления достоверности кластеризации является оценивание ее статистической значимости в смысле вероятности случайного опровержения нулевой гипотезы об отсутствии кластеризации. Верификация является важным элементом научных исследований. Отсутствие верификации или ее некорректное проведение может приводить к необоснованным и нередко ложным выводам. Данное утверждение, несомненно, должно относиться ко всем методам поиска закономерностей в данных, в том числе и к кластеризации.

Для того чтобы содержательно интерпретировать эффекты, связанные с кластеризацией, нужно с приемлемой степенью достоверности убедиться, что они существуют, что данные концентрируются в нескольких разделенных областях признакового пространства (т. е. что вероятность соответствия данных нулевой гипотезе, предполагающей генерацию всех данных из одного и того же равномерного или унимодального распределения, не имеющего соответственно кластерной структуры, мала, например менее 0.05). Описание нулевых гипотез, применяемых при валидации результатов кластеризации, можно найти, например, в [Gordon, 1996; Giancarlo, Utro, 2012].

Мерой статистической значимости предположения о реальном существовании кластеризации, полученной на реальных данных, может служить вероятность случайного достижения или превышения значения соответствующего индекса оценки качества кластеризации над значениями индексов оценки качества кластеризаций, полученных на данных, генерируемых при условии справедливости нулевой гипотезы.

В статистике такие вероятности принято называть *p*-значениями. Одним из способов оценивания *p*-значений является использование методов Монте-Карло, когда значения статистики критерия на реальных данных сравниваются со значениями статистики критерия на данных, сэмплированных из распределения в соответствии с нулевой гипотезой с использованием генераторов случайных чисел. Данный подход достаточно широко используется для верификации разнообразных регрессионных связей. Однако для верификации результатов кластерного анализа случайное сэмплирование используется, на наш взгляд, значительно реже, особенно в задачах с панельными данными. Вместе с тем такие задачи обладают существенной спецификой, связанной с неоднозначностью выбора нулевой гипотезы.

В частности, нулевая гипотеза может заключаться в том, что все рассматриваемые временные ряды для каждого показателя являются реализациями некоторого случайного процесса с одинаковыми характеристиками (средним, дисперсией, длиной ряда и т. д.). В этом случае временные ряды образуют фактически один кластер, а выявление большего числа кластеров является артефактом.

В данном исследовании демонстрируется методология верификации рядом альтернативных способов существования более чем одного кластера в данных на примере исследования совокупности регионов Российской Федерации в пространстве признаков, характеризующих их производственные функции. Выбор примера обусловлен тем, что ранее авторами подобная методология применялась для исследования производственных функций российских регионов [Кирилюк, Сенько, 2020].

В работах ряда исследователей, например в [Айвазян и др., 2016; Бахитова и др., 2014; Магомадов, Шамилев, 2014] и во многих других, приводятся варианты классификаций и кластеризаций регионов России по группам с использованием производственных функций, или наборов некоторых экономических показателей. В качестве примеров публикаций, где для этих задач используется иерархическая кластеризация, можно привести [Нижегородцев, Горидько, 2014; Сибукаев, 2019]. Интерес к этой теме обосновывается тем, что выявление достаточно достоверных кластеров позволяет более корректно проводить статистические расчеты, а также предполагает дальнейшие исследование по выявлению приведших к их возникновению механизмов, что может увеличить точность прогнозов. Для регионов из одного кластера могут быть полезны выработка единых рекомендаций, разработка общих программ сбалансированного развития.

Наш подход позволяет давать математическую оценку обоснованности разбиений регионов на кластеры, когда они производятся на основе наборов количественных признаков.

В публикациях, исследующих совокупность экономических объектов, на наш взгляд, могут быть выделены следующие подходы к использованию кластерного анализа:

1) объекты рассматриваются без учета их неоднородности;
2) проводится деление объектов на группы, но без применения кластерного анализа;
3) кластерный анализ проводится, но не производится оценка качества получившейся кластеризации с помощью соответствующих индексов;
4) качество кластеризации оценивается при помощи соответствующих индексов, но не решается задача оценки вероятности случайного возникновения высоких значений индексов (их возникновение можно рассматривать как ложную кластеризацию);
5) проводится кластерный анализ, вычисляются индексы качества кластеризации и оценивается вероятность случайного появления получившихся их значений.

Наш опыт говорит, что количество публикаций, которые можно сопоставить номерам пунктов приведенного перечня, существенно уменьшается с ростом номера.

2. Используемые данные

Нами произведена оценка кластерной структуры в пространстве показателей 79 регионов Российской Федерации, для которых есть нужный набор данных за рассматриваемый период [Регионы России…, 2017] (по данным за 1996–2014 гг.). Используемые панельные данные (те же, что ранее применялись нами в вышеупомянутой статье [Кирилюк, Сенько, 2020] для построения производственных функций регионов) включают следующие показатели: Y — валовый региональный продукт, I — инвестиции в основной капитал, L — среднегодовая численность занятых в экономике, помноженная на среднемесячную номинальную начисленную заработную плату работающих в экономике. Приведение величин к постоянным ценам (процедура, устраняющая искажающее влияние инфляции) осуществлялось с использованием индексов потребительских цен. Все используемые признаки были прологарифмированы. Средние значения временных рядов показателей, их дисперсии, тренды и прочие подобные характеристики образуют трехмерные пространства, что дает возможность наглядно оценить их кластерную структуру.

3. Методы кластеризации

Методы кластеризации делятся на неиерархические, типичным представителем которых является, например, метод *k*-средних, и иерархические. Авторами применяется агломеративная иерархическая кластеризация, когда в результате работы соответствующего алгоритма создается иерархия (дерево) вложенных кластеров. Преимущество иерархической кластеризации над альтернативными подходами заключается в том, что при ее применении не нужно выдвигать априорных предположений о числе кластеров.

Существует ряд метрик, характеризующих дистанции между временными рядами (например, dynamic time warping, манхэттонское расстояние и т. д.). В данной статье для данных, характеризующих объекты, практически синхронно развивающиеся во времени, и составляющих относительно короткие временные ряды, авторы предпочли использовать евклидову метрику. Используются два альтернативных подхода с применением евклидовой метрики.

1. Вычисление средних по времени значений рядов и последующее применение к ним алгоритма кластерного анализа. Кластеризация производится в пространстве из трех признаков, являющихся средними значениями по интервалу наблюдений показателей $Ln(Y)$, $Ln(I)$, $Ln(L)$. Расстояния между регионами i и j вычисляются в этом случае по формуле

$$d_{xij} = \left[\sum_{p=1}^{3} \left(\overline{x}_{jp} - \overline{x}_{ip} \right)^2 \right]^{1/2}, \tag{1}$$

где черта над x_{ip}, x_{jp} означает усреднение по времени значений показателей $Ln(Y)$ при $p = 1$, $Ln(I)$ при $p = 2$ и $Ln(L)$ при $p = 3$ для этих регионов.

2. Вычисление дистанций между трехкомпонентными временными рядами с использованием разностей их значений за все 19 лет рассматриваемого периода наблюдения и кластеризация с использованием этих дистанций:

$$\left[d_{xij} = \sum_{p=1}^{3} \sum_{t=1}^{19} \left(x_{jpt} - x_{ipt} \right)^2 \right]^{1/2}, \tag{2}$$

где x_{ipt}, x_{jpt} — значения признаков i-го и j-го регионов в год t.

Существует ряд альтернативных методов определения межкластерного расстояния при агломерации. В данной работе использовались три метода: complete, average, single, где расстояние между кластерами определяется соответственно как расстояние между наиболее удаленными друг от друга элементами двух кластеров, среднее расстояние между всеми парами элементов и расстояние между наиболее близкими элементами. Алгоритмы complete находят более компактные кластеры, а single, наоборот, — кластеры сложной формы, вытянутые, и они более чувствительны к шуму.

4. Индексы оценки качества кластеризации

Из множества существующих индексов оценки качества кластеризации нами отобраны для исследования (как одни из наиболее популярных) два: индекс Данна [Dunn, 1974] и Силуэт (Silhouette) [Rousseeuw, 1987].

Индекс Данна здесь используется в первоначальном варианте (существует ряд его модификаций) и определяется формулой

$$D = \min_{i,j \in \{1\ldots c\}, i \neq j} \left\{ \frac{d(c_i, c_j)}{\max_{k \in (1\ldots c)} \operatorname{diam}(c_k)} \right\}, \tag{3}$$

где d — расстояние между кластерами c_i, c_j; $\mathrm{diam}(c_k)$ — максимальное расстояние между элементами одного кластера.

Силуэт всей кластерной структуры (Silhouette Width Criterion — SWC) определяется по формуле

$$swc = \frac{1}{N_x}\sum_{j=1}^{N_x} S_{xj}, \tag{4}$$

как деленная на количество элементов в кластеризуемом множестве N_x сумма Силуэтов каждого отдельного элемента, определяемых по формуле

$$S_{xj} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})}, \tag{5}$$

где a_{pj} — среднее расстояние от объекта до других объектов своего кластера, b_{pj} — среднее расстояние от объекта до других объектов ближайшего другого кластера.

Для Силуэта (4)–(5), как и для индекса Данна (3), выполняется правило: чем выше качество кластеризации при данном числе кластеров, тем больше значение индекса.

5. Метод оценки статистической значимости кластеризации

Опишем применяемый в данной статье алгоритм оценки достоверности кластеризации посредством методов Монте-Карло. Генерируются псевдовыборки для имитации $Ln(Y)$, $Ln(I)$, $Ln(L)$ (независимо друг от друга). Они соответствуют двум вариантам нулевой гипотезы об отсутствии кластеризации, или, что то же самое, о существовании одного единственного кластера. Используется 2 варианта генерации псевдовыборок с длиной рядов, равной длине рядов используемых реальных данных (по 5000 псевдовыборок):

1) ряды, определяемые формулой

$$x_t = e_t, \tag{6}$$

где e_t — белый шум *iid* с нормальным распределением, длина рядов; средние значения и дисперсии реализаций процесса берутся равными усредненным по регионам значениям реальных исследуемых временных рядов признаков;

2) ряды, определяемые формулой

$$x_{t+1} = x_t + e_{t+1}, \tag{7}$$

которые имеют свойство стохастической нестационарности, могут демонстрировать эффект ложной регрессии [Granger, Newbold, 1974] и относятся к процессам случайного блуждания. Начальные значения для них генерируются из нормальных распределений со средними значениями и дисперсиями, равными усредненным средним значениям и дисперсиям по совокупностям реальных данных признаков регионов; длина рядов и дисперсии приращений равны соответствующим усредненным по регионам значениям для реальных рядов признаков; среднее значение приращений равно нулю. Случайные блуждания по результатам ряда исследований, например [Nelson, Plosser, 1982], значительно лучше описывают многие временные ряды экономических данных, чем стационарные процессы типа белого шума.

Полученные псевдовыборки, как и реальные данные, исследуются описанными выше методами оценки качества кластеризации.

Для каждого из описанных вариантов кластеризации в результате расчетов получены графики зависимости значений используемых индексов оценки качества кластеризации от предполагаемого числа кластеров.

На каждом графике зависимости индекса от числа кластеров откладывается по 7 типов данных: медианы значений индексов и границы их доверительных интервалов на уровне 5 % для симуляций белым шумом и случайными блужданиями, а также индексы, соответствующие реальным данным.

Достоверность кластеризации оценивается сравнением значений индексов реальных данных с соответствующими квантилями индексов из используемых псевдовыборок. Например, если индексы реальных данных больше по величине, чем индексы, соответствующие 95%-ным квантилям псевдовыборок, кластеризация принимается достоверной на уровне $p = 0.05$.

6. Результаты расчетов

6.1. Визуализация используемых данных

Поскольку используемое признаковое пространство имеет всего три измерения, кластерная структура в нем может быть легко оценена непосредственной визуализацией. На рис. 1 изображены три проекции набора значений показателей для всех регионов за все рассматриваемые годы (левые графики), а также три проекции набора значений показателей всех регионов, усредненных по используемому временному интервалу (правые графики).

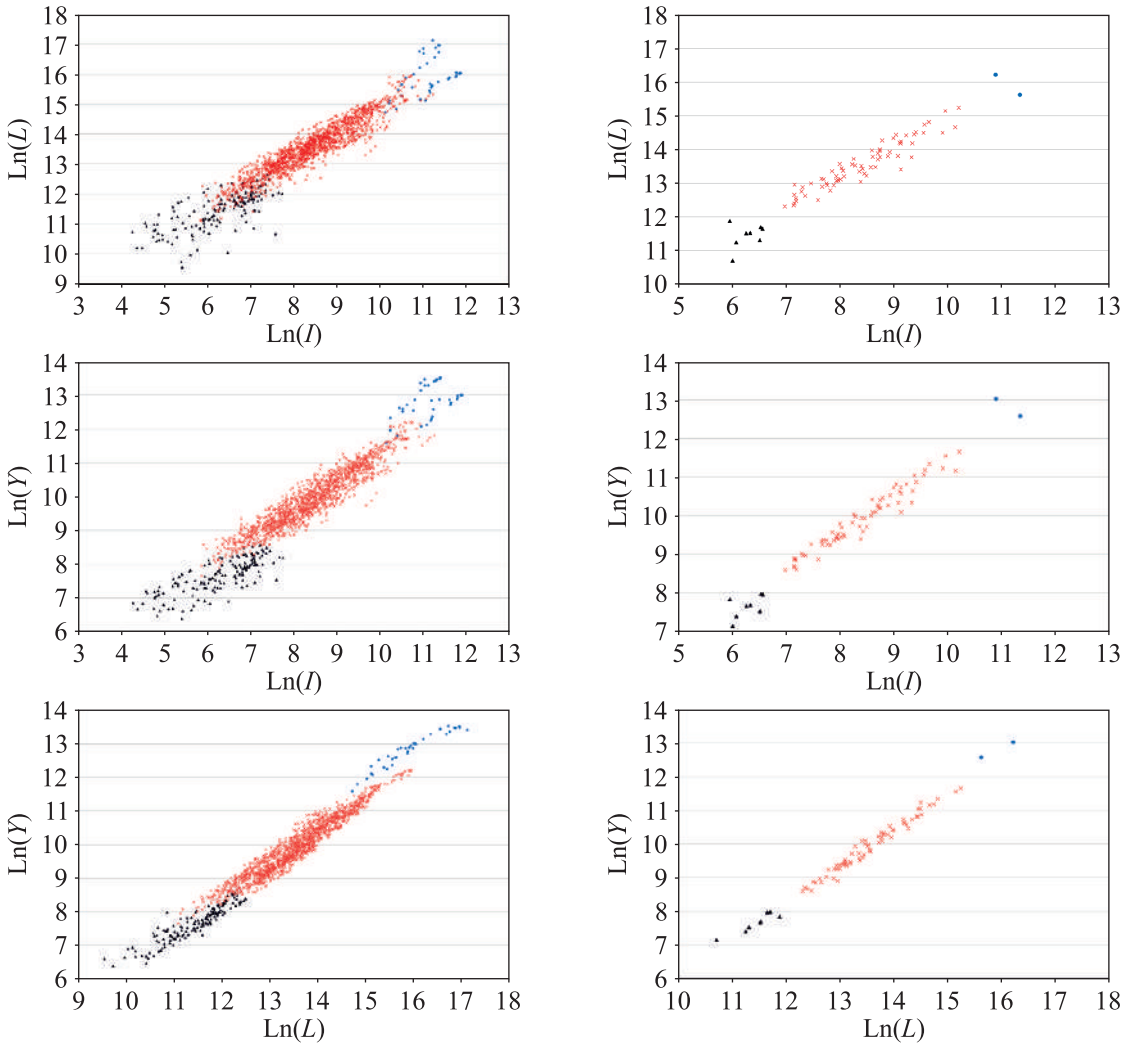

Рис. 1. Визуализация кластерной структуры российских регионов в пространстве признаков *Ln*(*Y*), *Ln*(*I*), *Ln*(*L*) с одним из вариантов разбиения на 3 кластера, обозначенных символами разных цветов

Как видно из рис. 1, данные расположены на всех графиках вдоль наклонных прямых (что обусловлено существенной корреляцией между рассматриваемыми признаками). При этом визуально можно выделить подгруппы точек, расположенных несколько обособленно от остальных и воспринимаемых субъективно как кластеры. Оценим качество кластеризации с помощью вычисления индексов реальных данных и сравнения их с индексами симуляций. Результаты этой оценки приводятся ниже.

6.2. Кластеризация по расстояниям, вычисленным по формуле (1)

На рис. 2 изображена зависимость значений индекса Данна от предполагаемого числа кластеров. Приводятся результаты только для методов average и complete, поскольку для метода single каких то качественно новых эффектов не выявлено.

На рис. 2 и на подобных рисунках ниже слева приведены результаты расчета методом average, справа — методом complete.

На рис. 2 и на последующих рисунках приняты следующие символьные обозначения:

▲ — 250-е по рангу (т. е. 95%-ные квантили), 2500-е по рангу (т. е. медианные) и 4750-е по рангу (т. е. 5%-ные квантили) значения индекса оценки качества кластеризации для симуляции по формуле (6);

× — 250-е по рангу, 2500-е по рангу и 4750-е по рангу значения индекса оценки качества кластеризации для симуляции по формуле (7);

● — реальные значения индекса.

Из рис. 2 видно, что значения индекса Данна, за исключением соответствующих самым малым N, возрастают для всех типов данных с ростом числа кластеров. Для метода average сильнее выражено нарушение монотонности роста индекса при малых N. Значения индекса для симуляций по формулам (6) и (7) почти сливаются на обоих графиках. Значения индекса для реальных данных нигде не расположены выше верхних границ доверительного интервала имитируемых. Однако в данном случае это не должно восприниматься как однозначное свидетельство того, что кластеризации данных нет. Для индекса Данна существуют примеры, когда он не различает четко выделяемые визуально, но слишком близко друг к другу расположенные кластеры.

На рис. 3 изображена зависимость значений индекса Силуэт от предполагаемого числа кластеров. Видны те же закономерности, что и для индекса Данна: постепенный рост значения индекса, за исключением самых малых N, слияние значений для симуляций по формулам (6) и (7). Отличием от результатов для индекса Данна является то, что Силуэт, особенно в случае применения метода average, показывает существенно лучшую достоверность кластеризации реальных данных в области малого числа кластеров.

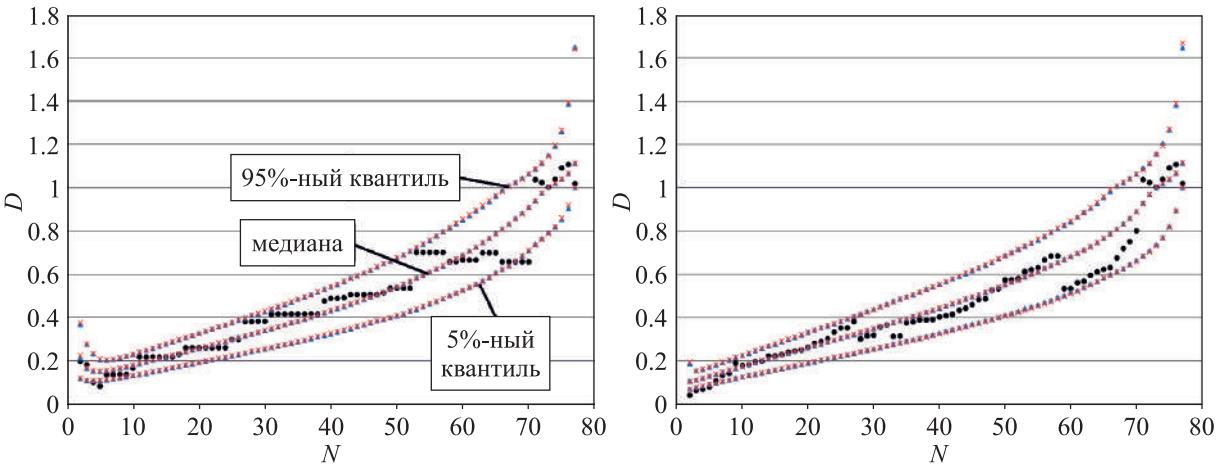

Рис. 2. Зависимость индекса Данна от числа выделенных кластеров для случая кластеризации с использованием дистанций (1)

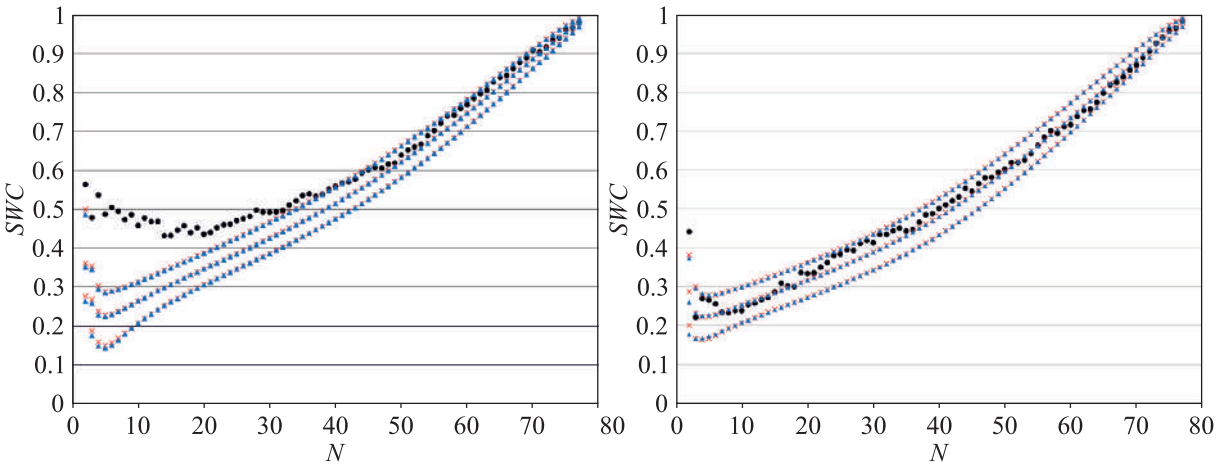

Рис. 3. Зависимость индекса Силуэт от числа выделенных кластеров при кластеризации с использованием дистанций (1)

6.3. Кластеризация по дистанциям, вычисленным по формуле (2)

В данном случае расстояния между реальными временными рядами существенно отличаются как от дистанций для симуляций (6), так и от дистанций для симуляций (7), что видно на рис. 4.

На рис. 4 изображены ранговые распределения расстояний между трехкомпонентными временными рядами, рассчитанных по формуле (2) для следующих видов данных (перечисляются в порядке расположения у оси ординат сверху вниз): реальные данные, симуляции по формуле (7), симуляции по формуле (6). Из рис. 4 видно, что распределение для реальных данных существенно отличается не только от распределения для симуляций (6), но и от распределения для симуляций (7), занимающих промежуточное положение по разбросу значений.

В отличие от графиков симуляций на рис. 2, 3 на рис. 5 значения индекса Данна для симуляций (6) и (7) четко разделены между собой. Из рис. 5 видно, что кривая значений индекса для реальных данных лежит намного ближе к кривым, полученным для симуляций (7), чем к кривым, полученным для симуляций (6), хотя в значительном числе случаев ниже 5%-ного квантиля симуляций (7).

На рис. 6 изображена зависимость значений индекса Силуэт от предполагаемого числа кластеров для варианта кластеризации с использованием дистанций (2). Оба графика демонстрируют достоверную кластеризацию при малых N. Неожиданным является то, что в отличие от случая, представленного на рис. 5, индексы для случайных блужданий расположены выше,

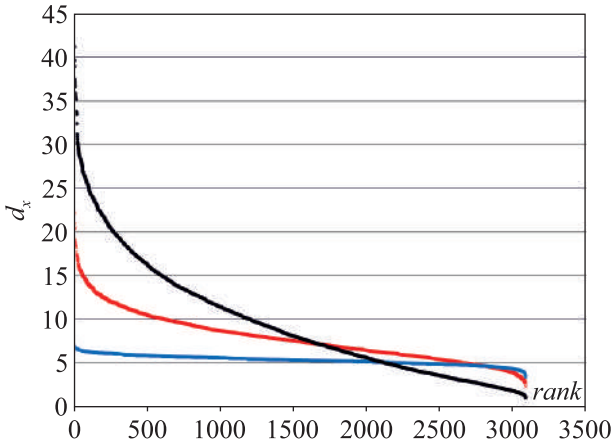

Рис. 4. Ранговые распределения дистанций d_x, вычисленных по формуле (2)

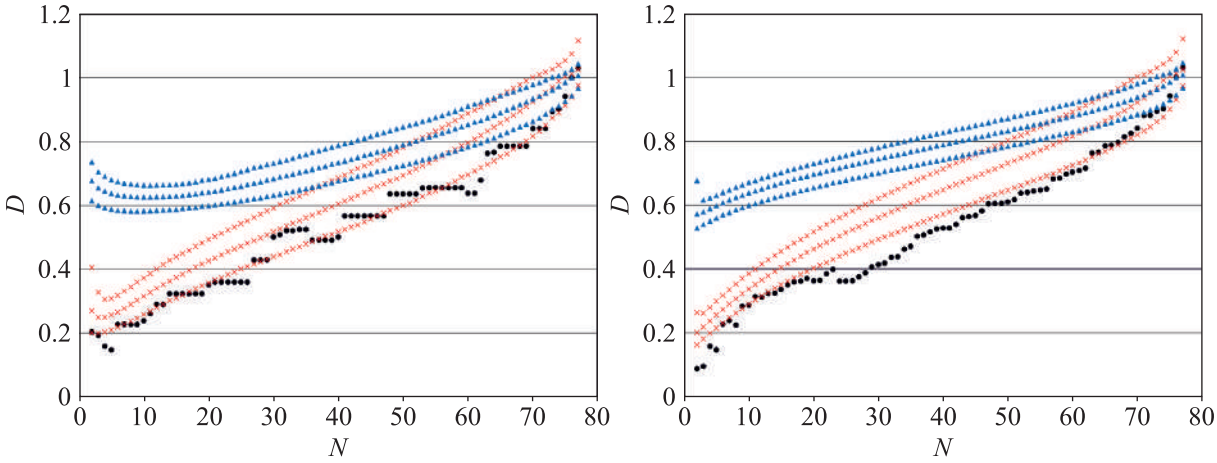

Рис. 5. Зависимость индекса Данна от числа выделенных кластеров при кластеризации с использованием дистанций (2)

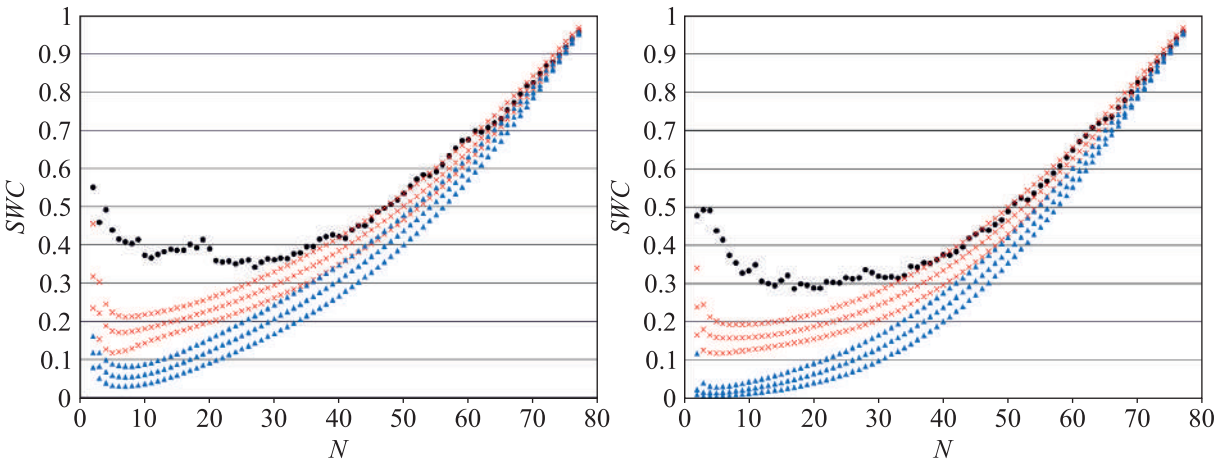

Рис. 6. Зависимость индекса Силуэт от числа выделенных кластеров при кластеризации с использованием дистанций (2)

чем индексы для белого шума. Это демонстрирует, что применение индексов качества кластеризации к слабо кластеризованным временным рядам и панельным данным требует определенной осторожности, поскольку поведение индексов в этом случае может иметь нетривиальные особенности, не имеющие отношения к их предназначению оценивать качество кластеризации.

6.4. Соотнесение регионов с кластерами

Для всех описанных выше вариантов кластеризации проверялось, как регионы распределяются по двум и по трем кластерам. Получены следующие результаты.

При делении на 2 кластера везде выделяются Москва и Тюменская область, которым соответствуют две относительно отдаленные от остальных точки в верхних левых частях графиков на рис. 1. Однако в случаях с использованием метода complete алгоритм добавляет в их кластер дополнительные регионы.

Для случая, рассчитанного по формуле (1), это Краснодарский край, Красноярский край, Московская область, Санкт-Петербург, Свердловская область, Республика Татарстан.

Для случая, рассчитанного по формуле (2), дополнительно к восьми регионам предыдущего случая в кластер добавляются Архангельская область, Республика Башкортостан, Челябинская область, Хабаровский край, Иркутская область, Республика Саха (Якутия), Кемеровская

область, Республика Коми, Ленинградская область, Нижегородская область, Новосибирская область, Омская область, Оренбургская область, Пермский край, Приморский край, Ростовская область, Сахалинская область, Самарская область, Саратовская область, Ставропольский край, Волгоградская область, Вологодская область, Воронежская область.

При делении на 3 кластера появляется кластер из регионов, находящихся в нижней левой части графиков на рис. 1. Он общий для всех использованных вариантов кластеризации. В него входят следующие регионы: Республика Адыгея, Республика Алтай, Чукотский автономный округ, Еврейская автономная область, Республика Ингушетия, Республика Калмыкия, Карачаево-Черкесская Республика, Республика Тыва.

Таким образом, результаты, полученные с использованием формул (1) и (2), различаются между собой только при применении метода complete. Метод complete выделяет 2 кластера существенно иначе, чем методы single и average.

Во всех случаях построение графиков визуально подтвердило адекватность кластеризации. То есть, даже когда достоверность кластеризации не подтверждаются индексами (в исследуемом нами случае — индексом Данна), распределение регионов по кластерам в основном может быть признано адекватным.

7. Заключение

В результате исследования качества агломеративной иерархической кластеризации панельных данных, характеризующих производственные процессы в регионах Российской Федерации, рядом альтернативных способов выявлено, что вычисляемая степень достоверности кластеризации существенно зависит от используемых индексов, выбора типа дистанций и методов вычисления расстояния между кластерами. Сформулируем наиболее значимые, по нашему мнению, выводы с рекомендациями для дальнейших исследований.

- При работе с панельными данными часто не учитывается возможность их кластерной структуры. Однако игнорирование кластерной структуры может приводить к существенным искажениям при эконометрическом моделировании.
- При обработке данных следует не только выяснять оптимальное разбиение их на кластеры, но и устанавливать его достоверность.
- Целесообразно использовать набор различных индексов оценки качества кластеризации, а не ограничиваться каким-то одним индексом, чтобы не прийти к ложным обобщениям в выводах.
- При выборе нулевых гипотез для проверки качества кластеризации временных рядов нужно учитывать факт их (не) стационарности, подбирать нулевые гипотезы об отсутствии кластеризации, соответствующие природе исследуемых временных рядов. Нестационарность временных рядов может приводить не только к ложной регрессии, исследуемой авторами, например, в [Кирилюк, Сенько, 2020], но и к ложной кластеризации. Незнание проблемы ложной регрессии [Granger, Newbold, 1974] до исследований К. Грэнджера приводило к бесчисленным фэйковым результатам. Использование дополнительных методов верификации позволило ограничить поток подобных «результатов». Учет возможности ложной кластеризации должен повысить научную значимость кластерного анализа как доказательного метода исследования. В настоящее время это преимущественно разведывательный метод.
- При этом Силуэт и индекс Данна дают разные ответы на вопрос о том, что обладает большей кластеризацией — набор реализаций белого шума (6) или случайных блужданий (7). Это свидетельствует, на наш взгляд, об определенной условности интуитивного понятия выраженности кластеризации в случаях, когда эта выраженность слаба. Для данных с сильнее выраженной кластеризацией, что соответствует, например, значительному превышению расстояний между кластерами над их характерными размерами, оба индекса в наших расчетах давали ожидаемые пики, соответствующие объективному числу кластеров.

- С помощью индексов кластеризации можно оценить не только качество кластеризации проверяемого набора временных рядов, но и степень его соответствия альтернативным нулевым гипотезам (например, сделать предположение о том, стационарны ли ряды).
- Исследованный набор данных по свойствам значимо отличается и от типичных реализаций белого шума (6) и от типичных реализаций случайных блужданий (7).
- Использование индекса Силуэт подтверждает достоверное наличие разбиения регионов на несколько кластеров, которое также можно оценить визуально. Поэтому, на наш взгляд, можно говорить о наличии кластерной структуры, хотя и не сильно выраженной, для рассматриваемой совокупности признаков, характеризующих производственные процессы в регионах Российской Федерации.

Представляет интерес продолжить исследования с использованием описанного в статье подхода с совместным применением большего числа индексов оценки качества кластеризации, других наборов признаков, в том числе многомерных, где проверка выраженности кластеризации посредством визуальной оценки затруднительна.

Все расчеты, результаты которых используются в данной статье, проведены с применением языка *R*, в частности пакетов *NbClust* [Charrad et al., 2014] и *TSclust* [Montero, Vilar, 2014].

Список литературы (References)

Айвазян С. А., Афанасьев М. Ю., Кудров А. В. Метод кластеризации регионов РФ с учетом отраслевой структуры ВРП // Прикладная эконометрика. — 2016. — Т. 41. — С. 24–46.
Aivazyan S. A., Afanas'ev M. Yu., Kudrov A. V. Metod klasterizatsii regionov RF s uchetom otraslevoi struktury VRP [The method of clustering regions of the Russian Federation taking into account the sectoral structure of the GRP] // Prikladnaya ekonometrika [Applied econometrics]. — 2016. — Vol. 41. — P. 24–46 (in Russian).

Бахитова Р. Х., Ахметшина Г. А., Лакман И. А. Панельное моделирование объема выпуска продукции для регионов России // Управление большими системами. — 2014. — Т. 50. — С. 99–109.
Bakhitova R. Kh., Akhmetshina G. A., Lakman I. A. Panel'noe modelirovanie ob"ema vypuska produktsii dlya regionov Rossii [Panel modeling of output for regions of Russia] // Upravlenie bol'shimi sistemami [Large-scale systems control]. —2014. — Vol. 50. — P. 99–109 (in Russian).

Ивахненко А. А., Каневский Д. Ю., Рудева А. В., Стрижов В. В. Выявление групп объектов, описанных набором многомерных временных рядов // Математические методы распознавания образов. — 2007. — Т. 13 (1). — С. 134–137.
Ivakhnenko A. A., Kanevskii D. Yu., Rudeva A. V., Strizhov V. V. Vyyavlenie grupp ob"ektov, opisannykh naborom mnogomernykh vremennykh ryadov [Identification of groups of objects described by a set of multidimensional time series] // Matematicheskie metody raspoznavaniya obrazov [Mathematical methods for pattern recognition]. — 2007. — Vol. 13 (1). — P. 134–137 (in Russian).

Кирилюк И. Л., Сенько О. В. Выбор моделей оптимальной сложности методами Монте-Карло (на примере моделей производственных функций регионов Российской Федерации) // Информатика и ее применения. — 2020. — Т. 14, вып. 2. — С. 111–118.
Kirilyuk I. L., Sen'ko O. V. Vybor modelei optimal'noi slozhnosti metodami Monte-Karlo (na primere modelei proizvodstvennykh funktsii regionov Rossiiskoi Federatsii) [Selection of optimal complexity models by methods of nonparametric statistics (on the example of production functionmodels of the regions of the Russian Federation)] // Informatika i ee primeneniya [Informatics and Applications]. — 2020. — Vol. 14, iss. 2. — P. 111–118 (in Russian).

Магомадов Н. С., Шамилев С. Р. Анализ динамики ВРП регионов РФ производственными функциями // Современные проблемы науки и образования. — 2014. — № 6.
Magomadov N. S., Shamilev S. R. Analiz dinamiki VRP regionov RF proizvodstvennymi funktsiyami [Analysis of the dynamics of GRP of the regions of the Russian Federation by production functions] // Sovremennye problemy nauki i obrazovaniya [Modern problems of science and education]. — 2014. — No. 6 (in Russian).

Нижегородцев Р. М., Горидько Н. П. Инновационные факторы экономического роста регионов России: кластерный анализ // Труды XII Всероссийского совещания по проблемам управления (ВСПУ-2014, Москва). — М.: ИПУ РАН, 2014. — С. 6088–6093.

Nizhegorodtsev R. M., Gorid'ko N. P. Innovatsionnye faktory ekonomicheskogo rosta regionov Rossii: klasternyi analiz [Innovative factors of economic growth in the regions of Russia: cluster analysis] // Trudy XII Vserossiiskogo soveshchaniya po problemam upravleniya (VSPU-2014, Moskva) [Proceedings of XII All-Russian Conference on Control Problems]. — Moscow: ICS RAS, 2014. — P. 6088–6093 (in Russian).

Регионы России. Социально-экономические показатели. 2017 // Стат. сб. / Росстат. — М., 2017.
Regiony Rossii. Sotsial'no-ekonomicheskie pokazateli [Regions of Russia. Socio-economic indicators]. 2017 // Stat. sb. / Rosstat. — Moscow, 2017 (in Russian).

Сибукаев Э. Ш. Изучение регионов России посредством иерархического метода кластерного анализа и данных о производстве // Университетская наука. — 2019. — № 2 (8). — С. 86–93.
Sibukaev E. Sh. Izuchenie regionov Rossii posredstvom ierarkhicheskogo metoda klasternogo analiza i dannykh o proizvodstve [Study of Russian regions by means of hierarchical method of cluster analysis and production data] // Universitetskaya nauka [University science]. — 2019. — No. 2 (8). — P. 86–93 (in Russian).

Сивоголовко Е. В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. — 2011. — № 4. — С. 14–31.
Sivogolovko E. V. Metody otsenki kachestva chetkoi klasterizatsii [Hard clustering validation methods] // Komp'yuternye instrumenty v obrazovanii [Computer tools in education]. — 2011. — No. 4. — P. 14–31 (in Russian).

Aghabozorgi S., Shirkhorshidi A. S., Wah T. Y. Time-series clustering — A decade review // Information Systems. — 2015. — Vol. 53. — P. 16–38. — https://doi.org/10.1016/j.is.2015.04.007

Charrad M., Ghazzali N., Boiteau V., Niknafs A. NbClust: An *R* Package for Determining the Relevant Number of Clusters in a Data Set // Journal of Statistical Software. — 2014. — Vol. 61, No. 6. — P. 1–36. — https://doi.org/10.18637/jss.v061.i06

Dunn J. Well Separated Clusters and Optimal Fuzzy Partitions // Journal Cybernetics. — 1974. — Vol. 4, No. 1. — P. 95–104. — https://doi.org/10.1080/01969727408546059

Giancarlo R., Utro F. Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis // Theoretical Computer Science. — 2012. — Vol. 428. — P. 58–79. — https://doi.org/10.1016/j.tcs.2012.01.024

Gordon A. D. Null Models in Cluster Validation // Gaul W., Pfeifer D. (eds) From Data to Knowledge. Studies in Classification, Data Analysis, and Knowledge Organization. — New York: Springer, 1996. — P. 32–44. — https://doi.org/10.1007/978-3-642-79999-0_3

Granger C. J., Newbold P. Spurious regressions in econometrics // Journal of Econometrics. — 1974. — Vol. 2. — P. 111–120. — https://doi.org/10.1002/9780470996249.ch27

Halkidi M., Batistakis I., Vazirgiannis M. On Clustering Validation Techniques // Journal of Intelligent Information Systems. — 2001. — Vol. 17, No. 2/3. — P. 107–145. — http://dx.doi.org/10.1023/A:1012801612483

Kapetanios G. Cluster analysis of panel data sets using non-standard optimisation of information criteria // Journal of Economic Dynamics and Control. — 2006. — Vol. 30, No. 8. — P. 1389–1408. — https://doi.org/10.1016/j.jedc.2005.05.010

Liao T. W. Clustering of time series data — a survey // Pattern Recognition. — 2005. — Vol. 38, No. 11. — P. 1857–1874. — https://doi.org/10.1016/j.patcog.2005.01.025

Montero P., Vilar J. A. TSclust: An R Package for Time Series Clustering // Journal of Statistical Software. — 2014. — Vol. 62, No. 1. — P. 1–43. — https://doi.org/10.18637/jss.v062.i01

Nelson Ch. R., Plosser C. I. Trends and random walks in macroeconomic time series: some evidence and implications // Journal of Monetary Economics. — 1982. — Vol. 10. — P. 139–162. — https://doi.org/10.1016/0304-3932(82)90012-5

Niu J. H. The Cluster Analysis of Multivariable Panel Data and Its Application // Applied Mechanics and Materials. — 2012. — Vols. 220–223. — P. 2668–2671. — https://doi.org/10.4028/www.scientific.net/amm.220-223.2668

Rousseeuw P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis // Journal of Computational and Applied Mathematics. — 1987. — Vol. 20. — P. 53–65. — https://doi.org/10.1016/0377-0427(87)90125-7