УДК 338.001.36
DOI: http://dx.doi.org/10.21686/2500-3925-2017-3-21-30

С.С. Шишулин
Финансовый Университет при Правительстве Российской Федерации, Москва, Россия

Методология сравнительного статистического анализа промышленности России на основе кластерного анализа

Статья посвящена изучению возможностей применения многомерного статистического анализа в исследовании промышленного производства России на основе сравнения его темпов роста и структуры с другими развитыми и развивающимися странами мира. Цель данной статьи заключается в определении оптимального набора статистических методов и последовательности их применения к данным промышленного производства, которые давали бы наилучший с точки зрения последующей содержательной интерпретации результат.

В качестве исследуемых данных выступают такие показатели структуры и динамики промышленного производства как индекс промышленного производства, выпуск, валовая добавленная стоимость, количество занятых и другие показатели системы национальных счетов и оперативной бизнес-статистики. Объектами наблюдения являются отрасли промышленного производства страны Евросоюза, Таможенного союза, США и Япония в 2005-2015 годах. В качестве инструмента исследования применены как простейшие приемы преобразований, графической и табличной визуализации данных, так и методы статистического анализа. В частности, на основе специализированного пакета программного обеспечения (СПСС) были применены метод главных компонент, дискриминантный анализ, иерархические методы кластерного анализа, метод Варда и к-средних.

Применение метода главных компонент к исходным данным позволяет существенно и эффективно сократить исходное пространство данных промышленного производства. Так, например, при анализе структуры промышленного производства сокращение составило с пятнадцати отраслей до трех основных, хорошо интерпретируемых, факторов: условно добывающие отрасли (с низкой степенью переработки), высокотехнологичные отрасли и отрасли товаров народного потребления (среднетехнологичные). При этом, в результате сравнения результатов применения кластерного анализа к исходным данным и данным, полученным на основе метода главных компонент, установлено что кластеризация данных промышленного производства на основе новых факторов значительно улучшает результаты кластеризации.

В результате анализа показателей разбиения данных на кластеры методами к-средних и иерархическими методами с использованием различных расстояний, было определено, что наилучший результат достигается при использовании комбинации данных методов, когда на первом этапе с помощью анализа визуализации иерархических алгоритмов (построения дендрограмм) определяется количество кластеров, на основе которого производится разбиение методом к-средних. При этом, значительное улучшение качества разбиения достигается за счет устранения в кластеризуемых данных выбросов, с последующим их включением в анализируемый набор с помощью дискриминантного анализа.

Применение данного подхода к данным структуры промышленного производства обеспечило высокие результаты. Полученные кластеры однородны по своему составу и содержательно интерпретируемы: в первый кластер входят страны с низкими показателями выпуска добывающей промышленности относительно совокупного выпуска экономики, при достаточно высоком значении данного показателя в других отраслях. В целом данную группу можно обозначить как страны с развитым промышленным производством высокотехнологичного типа. Вторая группа стран относительно других групп характеризуется в целом невысокой долей промышленности в экономике, и в частности более низкими показателями добывающих производств. К третьей группе стран относятся страны с высоко сырьевой базой, что характеризуется высокой долей в выпуске добывающих.

Ключевые слова: *промышленное производство, многомерный статистический анализ, сравнительный анализ, Россия, методология*

Sergey S. Shishulin
Financial University under the Government of Russian Federation, Moscow, Russia

Methodology comparative statistical analysis of Russian industry based on cluster analysis

The article is devoted to researching of the possibilities of applying multidimensional statistical analysis in the study of industrial production on the basis of comparing its growth rates and structure with other developed and developing countries of the world. The purpose of this article is to determine the optimal set of statistical methods and the results of their application to industrial production data, which would give the best access to the analysis of the result.

Data includes such indicators as output, output, gross value added, the number of employed and other indicators of the system of national accounts and operational business statistics. The objects of observation are the industry of the countrys of the Customs Union, the United States, Japan and Erope in 2005-2015. As the research tool used as the simplest methods of transformation, graphical and tabular visualization of data, and methods of statistical analysis. In particular, based on a specialized software package (SPSS), the main components method, discriminant analysis, hierarchical methods of cluster analysis, Ward's method and k-means were applied.

The application of the method of principal components to the initial data makes it possible to substantially and effectively reduce the initial space of industrial production data. Thus, for example, in analyzing the structure of industrial production, the reduction was from fifteen industries to three basic, well-interpreted factors: the relatively extractive industries (with a low degree of processing), high-tech industries and consumer goods (medium-technology) sectors. At the same time, as a result of comparison of the results of application of cluster analysis to the initial data and data obtained on the basis of the principal components method, it was established that clustering industrial production data on the basis of new factors significantly improves the results of clustering.

As a result of analyzing the parameters of data partitioning into clusters using k-means and hierarchical methods using different distances, it was determined that the best result is obtained when using a combination of these methods, when in the first stage the number of clusters is determined by analyzing the visualization of hierarchical algorithms (dendrogram construction) , On the basis of which the division by the method of k-means is made. At the same time, a significant improvement in the quality of the partition is achieved by eliminating the emissions in the clustered data, and then including them in the analyzed set using discriminant analysis.

The application of this approach to the data of the structure of industrial production ensured good results. The resulting clusters are uniform in composition and meaningfully interpreted: the first cluster includes countries

with low rates of output of the extractive industry relative to the cumulative output of the economy, with a sufficiently high value of this indicator in other sectors. In general, this group can be designated as a country with a developed industrial production of a high-tech type. The second group of countries with respect to other groups is characterized by a generally low share of industry in the economy, and in particular by lower rates of extractive industries. The third group of countries includes countries with a high resource base, which is characterized by a high share in the output of extractive industries.

Keywords: *industrial production, multidimensional statistical analysis, comparative analysis, Russia, methodology.*

Введение

В современном мире, при стремительном развитии непроизводственных сфер экономики, промышленное производство все еще остается основой экономического роста, ведь именно промышленность создает конечные потребляемые материальные блага, оказывающие прямое воздействие на уровень жизни населения.

Промышленное производство является главной, ведущей отраслью материального производства, в которой значительная часть валового внутреннего продукта и национального дохода. В современной экономике доля промышленности в совокупном ВВП развитых стран может доходить до 35%, в России по итогам 2016 года данный показатель составил 26,2% ВВП, из которых на долю обрабатывающего производства приходится 13,7 п.п., а добывающего 9,4 п.п. [1]

Ведущая роль промышленности обусловлена так же и тем, что от успехов в ее развитии зависит степень удовлетворения потребностей общества в высококачественной продукции, обеспечение технического перевооружения и интенсификации производства.

В связи с этим постоянный контроль тенденций развития промышленности является одним из ключевых направлений работы при определении уровня и тенденций развития экономики государства в целом. При этом, стоит отметить глобализацию современной экономики и особую зависимость промышленности России от спроса на внешних рынках, что делает необходимым анализ тенденций реального сектора отечественной экономики относительно общемировых или хотя бы обще европейских темпов и направлений экономического развития. Опережающие темпы роста или более низкие темпы падения соответствующих отраслей при этом можно отнести на результат промышленной политики государства. Это позволяет дать реальную оценку эффективности выполнения стратегии развития реального сектора экономики России. [2]

Основная идея сравнительного статистического анализа промышленного производства России заключается в том, чтобы давать оценку развитию, а именно, структуре и динамике промышленного производства, опираясь на краткосрочные и среднесрочные тенденции промышленного производства как части мировой экономической системы, без отрыва от нее. В первую очередь для этого необходимо определить эти тенденции, определить страны, имеющие те или иные признаки соответствующих тенденций, сопоставить экономики этих стран с экономикой России, в результате чего дать оценку, имеющую четкую систему координат относительно мирового уровня развития промышленного производства. Техническая реализация представленной идеи предполагает разбиение в каждый наблюдаемый период (2005–2015 года) стран на группы со схожими показателями структуры и динамики промышленного производства и последующее их исследование с помощью описательной статистики.

1. Подготовка исходных данных

Анализируемые данные, полученные на основе согласованной системы статистических показателей, хотя и являются согласованными на методологическом уровне, не могут использоваться без предварительной подготовки. [3] Главным образом это связанно с тем, что существующие международные стандарты в области статистики часто не имеют жестко регламентированных стандартов публикации и раскрытия соответствующих показателей. Выбор формы их публикации определяется национальными статистическими органами, которые, в свою очередь, ожидаемо решают задачу максимально точно отразить процессы характерные для данной экономики.

Первой проблемой является публикация данных в различном разбиении и группировке относительно видов экономической деятельности. Главенствующую роль здесь играет фактор международного разделения труда, так страны, ориентированные на производство высокотехнологичной продукции (например, Япония) более подробно раскрывают и публикуют информацию, относящуюся к производству машин, оборудования и прочей электроники, а страны, ориентированные на отрасли добывающей промышленности, имеют более подробную разбивку публикуемых показателей в этой области. Данная проблема решается путем составления агрегирующих и связывающих переходных шаблонов и таблиц на основе анализа используемых классификаторов. [4,5]

Второй проблемой использования обозначенных показателей является их публикация в различных формах – как в виде индексов, так и в виде абсолютных величин. При этом

абсолютные величины также могут не совпадать, например, данные могут быть представлены в штуках или тыс. штук, в долларах США или национальной валюте и так далее. Поэтому необходимо привести все имеющиеся показатели к единым формам.

Третьим аспектом использования анализируемых данных является наличие в них пропусков и работа с ними. Принципиальное значение наличия имеют показатели индекса промышленного производства и валовой добавленной стоимости, показатели статистики труда, цен и товарооборота имеют меньшее значение ввиду своей более слабой содержательной нагрузки в рамках анализа динамики и структуры промышленности, но имеют огромное значение в рамках определения и описания состояния промышленности. Поэтому исключение наблюдений производится только по первым двум параметрам. Критерием определения недоступности показателей структуры является сумма долей по наблюдению меньше 80%, по показателям структуры отсутствие более 3 показателей из 17 анализируемых.

Таким образом, для сравнительного статистического анализа структуры и динамики промышленного производства России использовался набор данных со следующими параметрами: период наблюдения 2005–2015 годы; 26–30 стран в зависимости от показателя; 15 отраслей для анализа структуры, 17 для анализа динамики; 6 показателей структуры, 5 показателей динамики; количество анализируемых значений 64 тыс.

2. Снижение размерности анализируемых данных

Исходные данные содержат в себе наблюдения, имеющие большое количество однотипных показателей, поэтому перед проведением кластерного анализа целесообразно попытаться снизить размерность анализируемых данных, выделив среди них главные компоненты, что, возможно, позволит улучшить результаты кластеризации, а так же определить группы взаимосвязанных отраслей, что представляет собой первый уровень анализа структуры промышленного производства. Основной проблемой при использовании данного метода является интерпретация полученных главных компонент т.к. они могут содержать в себе содержательно несвязные переменные. [6]

Для анализа данных промышленного производства воспользуемся реализацией метода главных компонент в специализированом программном обеспечении СПСС. Для примера, проанализируем полученные статистики по выделению главных компонент в структуре промышленности (табл. 1).

Таблица 1

Объясненная компонентами совокупная дисперсия

Компонент	Начальные собственные значения		
	Всего	% дисперсии	Суммарный %
1	3,7	25,1	25,1
2	2,3	15,8	41,0
3	1,8	12,2	53,2
4	1,6	11,2	64,5

Как следует из представленных выше данных, всего было выделено 4 компонента. Общий процент объясняемой ими дисперсии исходных данных составляет 64%, что является приемлемым результатом. Стоит отметить, что в данных отсутствует ярко выраженный фактор, что говорит о достаточной информативности анализируемого набора. Полученные компоненты можно считать равнозначными. Для содержательного анализа полученных компонент необходимо проанализировать матрицу нагрузок или корреляций полученных компонент с исходными данными (табл. 2).

Как видно из матрицы корреляций первый компонент имеет большие значения (0.7–0.8) с таки отраслями промышленности как машиностроение (C28), производство транспортных средств и прочей продукции связанной с транспортом (C29-C30), производство элек-

Таблица 2

Матрица нагрузок компонентов

Код эк. вида деятельности	Компонент			
	1	2	3	4
B	–0,157	–0,305	0,321	–0,544
C10-C11	–0,569	0,407	0,364	0,361
C13-C14	–0,097	0,830	–0,005	0,097
C16-C18	0,252	0,267	–0,589	–0,158
C19	0,073	0,141	0,758	–0,102
C20-C21	0,078	–0,236	0,050	0,843
C22-C23	0,420	0,761	0,034	0,076
C24	0,182	0,147	0,688	–0,236
C25	0,685	0,386	–0,434	–0,060
C28	0,745	–0,105	0,183	0,202
C26	0,552	–0,249	–0,245	0,284
C27	0,807	0,357	0,005	0,199
C29-C30	0,777	0,197	0,156	–0,035
C31-C32	0,089	0,061	–0,079	0,695
D	0,063	0,696	0,067	–0,159

трооборудования (С25), производство готовых металлических изделий (С25) и производство электроники (С26). При этом данный фактор имеет отрицательную корреляцию или низкую корреляцию с другими отраслями. Таким образом, содержательно данный фактор может трактоваться как уровень (или объём) тяжелого и высокотехнологичного машиностроения, называемый далее базовым машиностроением.

Второй компонент, напротив, имеет высокую корреляцию (0,7–0,8) с отраслями легкой (потребительской) промышленности, а именно с производством текстиля и одежды (С13-С14), производством изделий из пластика (резины, прочих неметаллических веществ). Кроме этого, данный показатель имеет высокую корреляцию (0,7) с производством тепла и электроэнергии, который также можно отнести к потребительски отраслям т.к. по сути это является коммунальными услугами. Этот же компонент имеет самую высокую корреляцию с производством пищевых продуктов (С10-С11). На основании всего этого, содержательной интерпретацией данного компонента является отрасли производства потребительских товаров.

Далее проанализируем четвертый компонент т.к. он имеет много общего со вторым. Во-первых, он имеет высокую (0,85) корреляцию с химическими (в т.ч. фармацевтическими) производствами и прочей промышленность (0,7) (С31-С32), к которой относятся производство мебели и украшений. Во-вторых, он также как и второй фактор имеет отрицательные корреляции с добывающими производствами. В целом данный фактор также содержательно можно отнести ко второй группе (потребительских) отраслей.

Таким образом, на одну содержательную единицу имеется два компонента. Для того чтобы снизить размерность необходимо объединить данные факторы в один содержательно интерпретируемый. Исходя из того, что по сути своей значения данных факторов представляют собой случайные величины, то их можно сложить, дисперсия нового показателя при этом также может быть посчитана простым сложением т.к. построение главных компонент предполагает, что они не коррелируют между собой. Таким образом, новый фактор будет нести в себе 28% дисперсии генеральной совокупности и становится равноценным по весу с первым фактором. Для простоты данный новый фактор обозначим как легкая промышленность.

Третий фактор имеет высокую корреляцию с добычей сырья и производствами его первичной переработки, к которым относятся производство кокса и нефтепродуктов (С19), производство металлов (С24) добывающее производство (В). Содержательно данный фактор имеет соответствующую интерпретацию, в дальнейшем для простоты именуемый как фактор добывающей промышленности.

3. Кластерный и дискриминантный анализ

На сегодняшний день существует большое разнообразие методов классификации объектов в т.ч. включающих в себя сложные алгоритмы и нейронные сети или требующие начального знания параметров целевых групп, наличия обучающих выборок. Несмотря на различные подходы к решению задач классификации, все методы основываются на представлении объектов в каком-либо пространстве, что делает их эффективными в работе с данными имеющими определенные параметры. При простой классификации объектов по нескольким числовым параметрам наиболее распространенными и часто встречающимися в специализированном программном обеспечении остаются иерархические статистические методы и метод к-средних, отличающиеся относительной простотой, высоким качеством получаемых результатов, их интерпретируемостью и широкими возможностями по настройке правил разбиения.

Иерархические методы классификации или кластерного анализа разделяются на два типа: агломеративные и дивизимные. Различие данных методов состоит лишь в том, что первые начинают алгоритм n элементов (классов) и далее объединяют близкие по расстоянию группы объектов, пока не останется всего один класс, а вторые наоборот начинают алгоритм с одного класса и разделяют дальние группы, пока не будет достигнуто разбиение n объектов на n классов. Исходя из описанного алгоритма, к несомненным преимуществам данного метода относится возможность построения дендограмм, т.е. деревьев, на которых четко видны этапы классификации и расстояние между классами. Основой работы алгоритма является матрица расстояний, которая формируется на основе правил объединений и расчета расстояний. Рассмотрим самые распространенные из них:

1) Метод ближайшего соседа. Расстояние между группами вычисляется как расстоянием между двумя максимально близкими точками (ближайшими соседями) этих групп.

2) Полная связь. Противоположен методу ближайшего соседа т.е. расстояния двумя группами принимается равным максимальному расстоянию между любыми двумя точками находящимися в различных группах.

3) Не взвешенное попарное среднее. Данный метод предполагает, что расстояние меж-

ду двумя различными группами определяется как среднее расстояние между всеми парами наблюдений находящихся в них.

4) Взвешенное попарное среднее. Аналогичен предыдущему методу, кроме того, что в качестве весов к определяемым расстояниям берутся размеры соответствующих групп, т.е. количество наблюдений в них.

5) Не взвешенный центроидный метод. Данный метод предполагает вычисление расстояния между двумя группами на основе их центров тяжести.

6) Взвешенный центроидный метод. Аналогичен предыдущему методу, кроме того, что в качестве веса к определяемым расстояниям берутся размеры соответствующих групп, т.е. количество наблюдений в них.

7) Метод Варда. Данный метод стоит отдельно от всех выше указанных, т.к. его алгоритм для определения расстояний между группами применяет методы дисперсионного анализа. Алгоритм на каждом этапе минимизирует сумму квадратов для любых двух возможных групп наблюдений. Это происходит следующим образом: в каждом кластере вычисляются центры; далее определяются и суммируются все расстояния от центра кластеров до наблюдений в них входящих; в новый кластер сливаются те наблюдения, при реализации которого при пересчете суммы расстояний её прирост будем минимален. [7,8]

Алгоритм классификации наблюдений методом к-средних, во многом похож с алгоритмом метода Варда, однако на последнем этапе он не предполагает объединения целых кластеров, а осуществляется миграция наблюдений. При этом если ограничением работы метода Варда является объединение на последнем этапе всех наблюдений в одну группу, то для работы метода к-средних необходимо ограничение в виде первоначально заданного числа кластеров, что требует априорного наличия информации об исследуемой совокупности и является одним из недостатков метода. При этом от выбора этих точек также зависит и конечный результат, в общем случае алгоритм гарантирует нахождение только локального минимума суммарного квадратичного отклонения. Метод включает в себя следующие этапы:

1) определяется число кластеров будущего разбиения;

2) в общем случае, случайно выбираются k наблюдений, значения параметров которых признаются начальными центрами кластеров;

3) для каждой точки набора данных вычисляется ближайший к ней центр, ближайшие точки группы объединяются;

4) определяются новые центры образованных кластеров, и повторяется шаг 3;

5) алгоритм прекращает работу, когда на n-ой итерации не происходит изменения центров имеющихся кластеров. [9]

Подводя итог, можно сказать, что иерархические методы достаточно просты для реализации, что одновременно является, как плюсом, так и их минусом. Не все описанные методы одинокого хороши для всех форм наблюдаемых данных. Например, метод ближайшего соседа склонен формировать кластеры в виде цепей, а полной связи рощ, метод Варда, хотя и очень эффективен, часто разделяет совокупность на неприемлемо большое количество групп. Метод к-средних схож с методом Варда, его результат более устойчив, однако он требует априорных знаний о выборке. Таким образом, применение какого-либо метода к неподходящим данным и последующая интерпретация полученного результата может привести к ошибочным выводам. Поэтому выбор метода должен быть обоснован предварительным сравнительным анализом всех возможных вариантов.

Применим изложенные выше подходы к кластеризации одного набора данных: данных структуры промышленного производства по выпуску за последний доступный в полном объеме период данных 2014 год. Для наглядного сравнительного анализа предложенных методов результат представлен в виде таблицы (табл. 3).

В таблицу записаны результаты кластеризации, при этом под оптимальным количеством кластеров понимается экспертно определенное их количес-

Таблица 3

Наилучшие результаты кластеризации

Правило разбиения	Исходные данные			Выделенные компоненты		
	Оптимальное разбиение	Пороговое расстояние / Количество наблюдений в кластерах	Количество выбросов	Оптимальное разбиение	Пороговое расстояние / Количество наблюдений в кластерах	Количество выбросов
Межгрупповой связи	2	95,6/5-26	3	3	3,5/6-8-17	3
Внутригруп. связи	2	42,5/5-19	10	4	2,1/5-5-8-16	2
Ближнего соседа	2	27,1/4-22	8	2	1,4/4-25	5
Дальнего соседа	4	109,1/3-5-4-15	7	3	7,4/10-5-19	нет
Не взвешенный центроид	2	92,8/3-28	3	4	2,2/5-8-5-12	4
Медианный	3	88,6/3-3-21	7	3	3,1/10-5-19	нет
Метод Варда	3	716/8-7-14	5	3	25,8/9-5-20	нет
Метод к-средних	3	Нет /2-6-32	2	3	Нет/16-9-9	нет
Метод к-средних	2	Нет /2 - 32	2	2	Нет/24-10	нет

тво на основе оценки дендрограммы и расстояний между кластерами: на соотношении приращения количества наблюдений в кластере и порогового расстояния т.е. расстояния на котором происходит объединение в кластер. При этом сравнивать показатели порогового значения возможно только в рамках одного набора данных т.к. его значение будет естественно больше в наборе исходных данных, где в несколько раз больше исходных наблюдений, а, следовательно, и большие расстояния между объектами. Тоже относится и к методу Варда, т.к. он предполагает другую меру расстояния. Под выбросами понимаются наблюдения, не вошедшие в оптимальные кластеры. Для метода к-средних понятия порогового расстояния и выброса отсутствует по определению, косвенно о качестве разбиения может служить количество наблюдений в итоговых кластерах, которые и записаны в таблицу.

Первое что стоит отметить, на основе сравнительного анализа результатов классификации данных по структуре промышленного производства различными методами кластеризации, это значительно превосходящее качество разбиения наблюдений на основе главных компонент. Об этом свидетельствуют, как и значительное снижение выбросов, так и качество распределения наблюдений по кластерам с точки зрения количественных характеристик.

Вторым выводом является утверждение о близости качества разбиения метода Варда и к-средних, о чем свидетельствуют их относительно одинаковые результаты разбиения, при этом метод к-средних распределяет наблюдения по кластерам более равномерно, что, несомненно, является его преимуществом. Составы кластеров практически не отличаются, например, третий кластер совпадают на 100%, однако некоторые существенные отличия все-таки имеются. Так, например, классификация по методу Варада отнесла Австрию, Швецию и Финляндию, которые относятся к странам с достаточно развитой и большой долей промышленности в экономике, к первой группе стран, где преобладают страны или с неразвитой промышленностью, или с развитой, но не обладающей большой доле из-за наличия других более значительных секторов экономики, например, финансового.

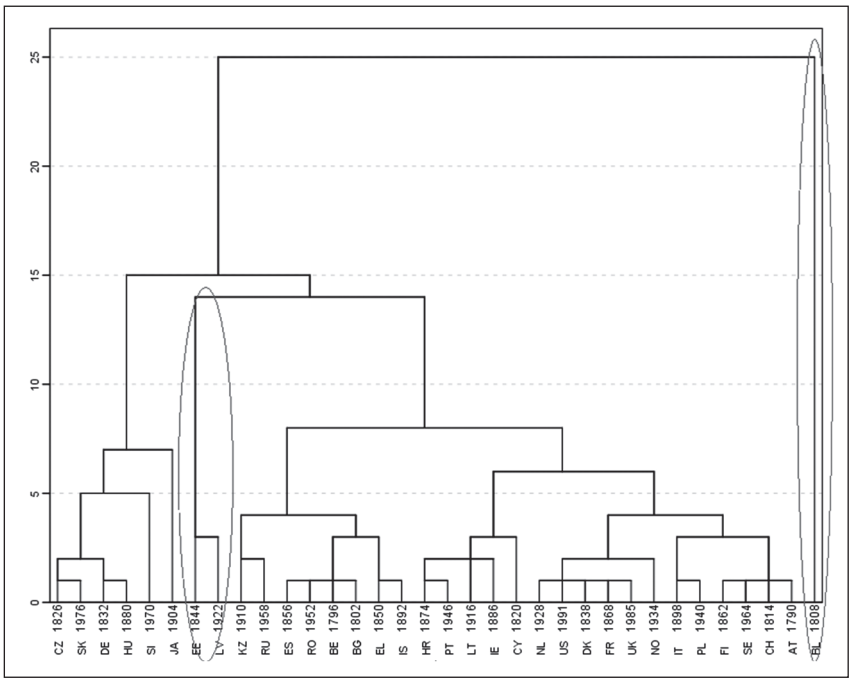

Рис. 1. Дендрограмма разбиения данных структуры промышленного производства с использованием метода межгрупповой связи

Таким образом, метод к-средних признается более оптимальным по сравнению с методом Варда как по количественной, так и по качественной структуре разбиения. Кроме этого, результат разбиения в отношении количества кластеров в данном методе строго детерминирован, что является с одной стороны положительным моментом т.к. позволяет анализировать устойчивые и сравнительно однородные группы во времени, а с другой стороны проблемой т.к. изначально данное количество необходимо определить.

Для решения этой проблемы на первом этапе, возможно, воспользоваться иерархическими методами, которые с достаточной точностью определяют количество кластеров. При этом данный метод позволяет строить дендрограммы, на основании которых легко определить аномальные наблюдения, которые влияют на качество итогового разбиения. Под аномальными наблюдениями подразумеваются наблюдения, значения показателей которых резко отличаются от имеющейся совокупности. Данные наблюдения легко определить путем анализа дендрограммы разбиения, на которой отражается расстояние между данным наблюдением и кластером или группой, в который оно включено на n-ом шаге (рис. 1).

На дендрограмме прослеживается наличие нескольких групп рядом расположенных наблюдений, которые объединяются в кластеры на небольшом расстоянии примерно в 5–8 единиц нормированной шкалы (левая шкала). При этом также наблюдается единственное наблюдение (Белоруссия), которое объединяется

со всей основной группой на расстоянии 25 единиц. Данное расстояние свидетельствует о том, что в принципе данное наблюдение было бы равнозначно включить в любую из групп, т.к. имеет крайне аномальные значения показателей в имеющейся мере расстояний. При этом при включении данного наблюдения в какой-либо кластер, оно серьезно будет смещать его центр, приводя к искажению классификации. Поэтому такого рода наблюдения необходимо исключать из анализируемых данных, при этом это не означает, что его полное исключение из анализа. После разбиения отфильтрованных данных, такие наблюдения можно «безболезненно» отнести к наиболее подходящему кластеру на основе дискриминантного анализа.

Дискриминантный анализ относится к тому же разделу статистики, что и кластерный, к многомерному статистическому анализу, однако относится к группе методов классификации объектов на основе максимального сходства при наличии обучающих параметров. Сущность дискриминантного анализа заключается в формулировке правила, на основе которого классифицируемым наблюдениям присваивается один из уже имеющихся обучающих кластеров. Данное правило реализуется на основе дискриминантной функции, значение которой для исследуемого объекта, вычисленное по его признакам, сравнивается с рассчитанными на основе обучающих выборок значениями дискриминации.[10]

Для описываемого случая, при кластеризации объектов без аномальных наблюдений (Белоруссия, Латвия и Эстони) имеются следующие результаты (табл. 4).

На основе полученной обучающей совокупности кластеров проводится дискриминантный анализ имеющихся исключенных наблюдений, для чего алгоритмом СПСС строятся две дискриминантные функции, основные показатели которых представлены в табл. 5.

Полученные функции имеют высокую разделительную способность, о чем свидетельствуют большие значения показателя канонической корреляции 0,83 и 0,81 соответственно. Относительный вклад каждой анализируемой переменной в значение функции хорошо виден из структурной матрицы, которая показывает, как отдельные коэффициенты коррелируют со значением дискриминантной функции. Для определения значения функции коэффициенты перемножаются на соответствующие исходные показатели анализируемых наблюдений. Для последующего определения того, к какому кластеру относится дискриминируемое наблю-

Таблица 4

Результаты кластеризации по данным с исключенными аномальными наблюдениями

Показатели	1 кластер		2 кластер		3 кластер	
	Среднее значение	Количество наблюдений	Среднее значение	Количество наблюдений	Среднее значение	Количество наблюдений
Базовая промышленность	–0,32	17	–0,70	8	1,72	6
Добывающая промышленность	–0,36		1,07		0,23	
Потребительская промышленность	–0,08		–0,07		0,083	

Таблица 6

Результаты дискриминантного анализа

№	Показатели	Функция		Коэффициент канонической корреляции	Значение функции в центроидах групп	
		Коэффициенты	Матрица структуры			
11	Базовая промышленность	0,918	0,823	0,829	1кл	–1,00
	Добывающая промышленность	0,540	0,110		2кл	0,09
	Потребительская промышленность	0,427	0,366		3кл	2,71
22	Базовая промышленность	–0,427	–0,497	0,810	1кл	–0,74
	Добывающая промышленность	0,908	–0,038		2кл	2,22
	Потребительская промышленность	0,173	0,874		3кл	–0,86

Таблица 6

Результаты дискриминации исключенных наблюдений

Страна	Базовая пром.	Добывающая пром.	Потреб. пром.	Значение функц. №1	Значение функц. №2	Кластер
Белоруссия	–0,22	2,96	1,00	4,05	5,9	2
Эстония	0,145	–2,26	0,133	–2,24	–4,52	1
Латвия	–0,98	–1,96	–0,038	–3,6	–3,28	1

дение, используются центры кластеров в терминах дискриминантной функции.

Рассмотрим имеющиеся аномальные наблюдения и подвергнем их дискриминантному анализу (табл. 6):

На основании дискриминантного анализа Белоруссия была отнесена ко второму кластеру, а Эстония и Латвия к первому. При этом стоит отметить, что полученный итоговый результат по своему качественному и количественному содержанию кластеров практически совпадает с классификацией методом Варда без исключения аномальных наблюдений. Во-первых это говорит о том, что исключение аномальных наблюдений значительно улучшают качество кластерного анализа, а во-вторых, что для определения количества кластеров и аномальных наблюдений предпочтительней использовать метод Варда, как изначально более эффективный. Стоит также отметить, что аномальные наблюдения были отнесены к одинаковым группам как методом Варда, так и дискриминантным анализом. При этом это не отменяет того факта, что аномальные наблюдения способны значительно исказить проводимое исследование и их не надо исключать при использовании метода Варда.

4. Интерпретация полученных результатов

Полученные в результате разбиения кластеры имеют хорошую содержательную интерпретацию. В первую группу входят страны с низкими показателями выпуска добывающей промышленности (0,4%) относительно совокупного выпуска экономики. При этом выпуск отраслей обработки сырь находится на достаточно высоком уровне – производства из дерева составляют 2,6% всего выпуска, металлургии 2,1%, а нефтепереработки 1,6%. Характерной чертой рассматриваемой группы также является высокие доли в выпуске отраслей автомобилестроения (4,8%), машиностроения (3%), электроприборостроения (1,8% и 2,8%), обработки металлов (2,4%). Основная часть валовой добавленной стоимости промышленности в данных странах генерируется за счет как раз этих отраслей, суммарно данный показатель составляет 9,6% вдс по экономике или 42,5% от вдс промышленности. Всего промышленность формирует 22,7% вдс всей экономики. В целом данную группу можно обозначить как страны с развитым промышленным производством высокотехнологичного типа.

Вторая группа стран относительно других групп характеризуется невысокой долей добывающего производства (1,1%) в выпуске, а также производства нефтехимии (0,9%) и металлов (0,7%). Очевидно, что это страны не только с низкими запасами сырья, но и отсутствием отраслей его переработки. Выпуск пищевых производств (4,7%), текстильных (0,9%), химических (2,6%) находится на уровне других стран. Стоит отметить, что данные отрасли равномерно развиты во всех экономиках. У данной группы нет явного преимущества в выпуске относительно обоих конкурирующих кластеров, при этом у данной группы по сравнению с третьей наблюдаются более высокие значения выпуска по отраслям деревопереработки (2%) и производства электроники (1,1%). Валовая добавленная стоимость, формируемая за счет промышленности, в среднем по группе составляет 16,3% от всей экономики, доля высокотехнологичных производств в ней при этом составляет всего 22,2%. В целом вторую группу стран можно охарактеризовать

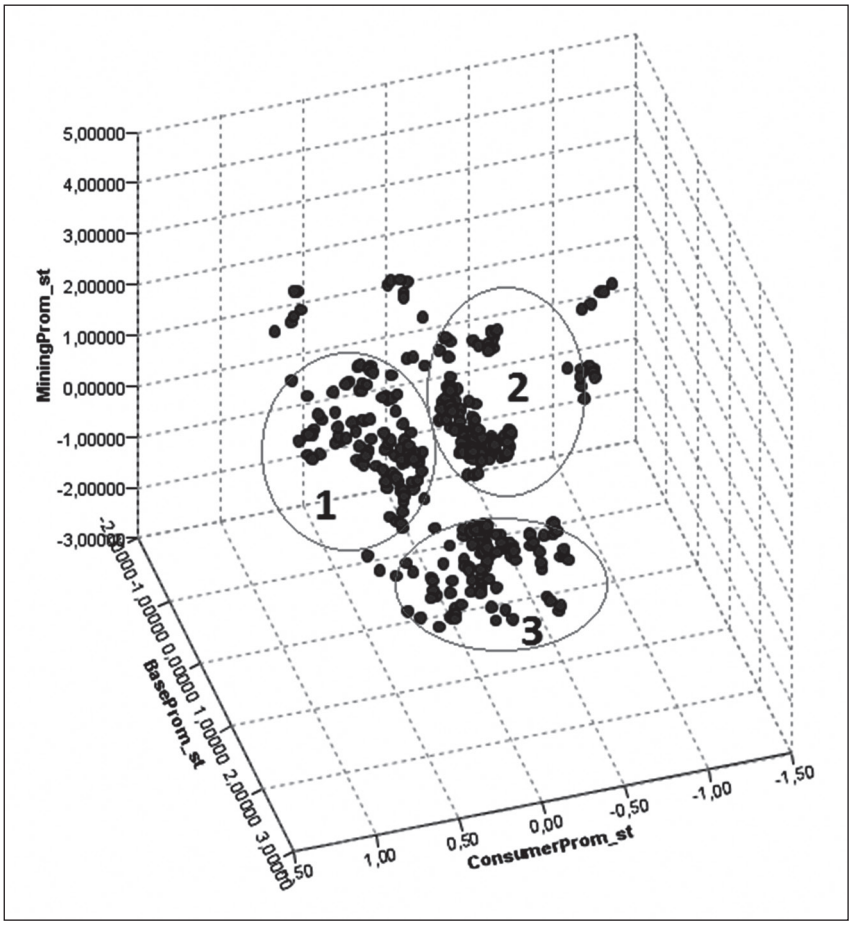

Рис. 2. Графический анализ структуры промышленного производства в пространстве выделенных главных компонент

как имеющие экономику с низкой ресурсной базой и низкой степенью развития высокотехнологичных отраслей обрабатывающей промышленности.

Из описания первой и второй групп становится ясно, что к третьей группе стран относятся страны с высоко сырьевой базой, что характеризуется крайне высокой долей в выпуске добывающих производств (4,2%), производства нефтепродуктов (3,3%) и металлов (2,9%), и невысоко промышленного высокотехнологичных производств: автомобилестроение (1,9%), машиностроения (1,4%), электроприборостроения (0,8% и 0,7%), обработки металлов (1,2%) и дереаообработки (1,3%). Всего промышленность данных стран формирует 22,6% валовой добелённой стоимости всей экономики, что сопоставимо с показателем первой группы. При этом высокотехнологичная промышленость составляет всего 4% вдс экономики или 17,9% вдс промышленности, что меньше показателя второй группы.

О качестве разбиения можно судить и по средствам графического анализа построенного в пространстве выделенных главных компонент графика (рис. 2). На рисунке явно выделяются три группы наблюдений или кластера, что также подтверждает целесообразность снижения размерности данным методом.

Заключение

Таким образом, исходя их проведенного анализа применения методов классификации к показателям промышленного производства, предлагается следующая методика кластеризации стран на основе показателей промышленного производства:

Анализ структуры анализируемых данных и выделение главных компонент;

Формирование содержательных факторов на основе главных компонент;

Определение на основе содержательных факторов и метода классификации Варда оптимального количества кластеров и аномальных наблюдений в каждом наборе исследуемых данных;

Кластерный анализ каждого набора отфильтрованных данных методом к-средних с количеством кластеров определенным на предыдущем шаге;

Дискриминантный анализ исключенных на втором шаге аномальных наблюдений.

В целом, работу алгоритма стоит признать удовлетворительной и пригодной для использования в сравнительном статистическом анализе промышленного производства.

К недостаткам предложенного алгоритма можно отнести большую трудоемкость его реализации даже с учетом снижения пространства за счет выделения главных факторов. В частности, для анализа промышленного производства, чтобы составить динамику изменения показателя необходимо проанализировать отдельный набор данных для каждого показателя в каждом году, что соответствует 20 повторениям только по показателям структуры по выпуску к выпуску по экономике и динамике физического объёма выпуска.

Кроме этого, полученные показатели классификации объектов во времени необходимо интегрировать в созданную систему показателей промышленного производства. Данная операция необходима для расчета как новых показателей (например, товарооборота со странами отдельных групп), так и для составления описательной статистики уже имеющихся.

В целом, подводя итог исследованию проблемы сравнительного статистического анализа промышленного производства России, можно сформулировать следующие основные этапы и положения методологии его реализации:

1) Определение анализируемой совокупности объектов и признаков наблюдения на основе существующих экономических связей;

2) Выбор основы и создание системы взаимосвязанных и согласованных показателе характеризующих анализируемую совокупность;

3) Подготовка и обработка исходных данных;

4) Классификация объектов наблюдения на основе составленной на втором шаге системы показателей и методики кластерного анализа стран на основе показателей промышленного производства;

5) Интеграция полученных значений классификации наблюдений в систему показателей, расчет показателей описательной статистики промышленного производства;

6) Оценка и интерпретация полученных значений для объекта исследования.

Литература

1. Социально-экономическое положение России (2016 год). Режим доступа: (Дата обращения: 05.05.2017)

2. *Шишулин С.С.* Сравнительный анализ темпов развития промышленного производства России и Евросоюза. Экономические науки. – 2015. – №8 (129)-2015 август – С. 99–103.

3. *Шишулин С.С.* Система экономико-статистических показателей структуры и динамики промышленного производства России. Мир новой экономики. – 2016 –Т. 10. №4. – С. 135–141

4. Statistical classification of economic activities in the European Community. NACERev. 2. Режим доступа: http://ec.europa.eu/competition/mergers/cases/index/nace_all.html (Дата обращения: 30.01.2017)

5. Общероссийский классификатор видов экономической деятельности (ОКВЭД2). Режим доступа:оквэд.рф/ (Дата обращения: 30.01.2017)

6. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики. – Т.1: Теория вероятностей и прикладная статистика М.: ЮНИТИ-ДАНА, 2001. – 656 с.

7. *Берндт Э.* Практика эконометрики. Классика и современность. М.: Юнити, 2005. – 863 с.

8. *Greene W.* Econometric Analysis. 7th Edition. – Prentice Hall, 2011. – 1231 p.

9. *Hayashi, F.* Econometrics. Princeton University Press, Princeton, 2000. – 683 p.

10. *Вербик М.* Путеводитель по современной эконометрике. М.: 2008. – 616 с.

References

1. Socialno-ekonimicheskor polozhenie Rossii (2016 god). [Electronic resource]: Available at: (Accessed: 05.05.2017) (in Russ.)

2. *Shishulin S.S.* Sravnitelniy analyz tempov razvitiya promishlennogo proizvodstca Rossii I Evrosouza. Ekonimicheskie nauki. – 2015 – №8 (129)-2015 avgust – p. 99–103 (in Russ.)

3. *Shishulin S.S.* Sistema ekonomiko-statisticheskih pokazateley strukturi I dinamiki promishlennogo proizvodstva Rossii. Mir novoy ekonomiki. – 2016 – T. 10. №4. – p. 135–141.

4. Statistical classification of economic activities in the European Community. NACERev. 2. [Electronic resource]: Available at: http://ec.europa.eu/competition/mergers/cases/index/nace_all.html (Accessed: 30.01.2017)

5. Obsherossiyskiy klassifikator vidov ekonomicheskoy deyatelnosti (OKVED2). (Accessed: 30.01.2017) (in Russ.)

6. Ayvozyan S.A., Mkhitoryan V.S. Prikladnaya statistica. Osnovi ekonometriki. – T.1: Teoriya veroyatnostey I prikladnaya statistica. M.: Unity-Dana, 2001. – 656 s.

7. *Berndt E.* Practika ekonometriki. Classika i sovremennost. M.: Unity, 2005. – 863 s. (In Russ.)

8. *Greene W.* Econometric Analysis. 7th Edition. – Prentice Hall, 2011. – 1231 p.

9. *Hayashi, F.* Econometrics. Princeton University Press, Princeton, 2000. – 683 p.

10. Verbik M. Putevoditel po sovremennoy ekonometrike. M.:2008. – 616 s. (In Russ.)

Сведения об авторе

Сергей Сергеевич Шишулин
Аспирант кафедры «Статистика»
Финансовый Университет при Правительстве Российской Федерации, Москва, Россия
Эл. почта: shishulinsergey@mail.ru

Information about the author

Sergey Sergeevich Shishulin,
graduate student, statistics department
Financial University under the Government of Russian Federation, Moscow, Russia
E-mail: shishulinsergey@mail.ru