

А.С. АХРЕМЕНКО, А.Л. МЯЧИН*

**ПАТТЕРН-АНАЛИЗ И КЛАСТЕРИЗАЦИЯ
В ИССЛЕДОВАНИИ ГОСУДАРСТВЕННОЙ
СОСТОЯТЕЛЬНОСТИ: «АДАПТИВНАЯ ОПТИКА»
ДЛЯ ПОЛИТИЧЕСКОЙ НАУКИ¹**

Аннотация. Центральный фокус работы – методологический. На примере набора индикаторов государственной состоятельности авторы показывают конкретную стратегию выявления устойчивых структур в многомерных массивах данных, отражающих сложные и неоднозначные понятия политической науки. Ключевая особенность этой стратегии – применение родственных, но существенно различающихся по своим техническим особенностям многомерных методов – кластерного и паттерн-анализа. В статье использована иерархическая кластеризация с различными сочетаниями метрик и правил объединения, а также порядково-инвариантная паттерн-кластеризация. Попутно впервые в политологической ли-

* **Ахременко Андрей Сергеевич**, доктор политических наук, доцент, замдекана по науке факультета социальных наук Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ, Москва, Россия), профессор Департамента политической науки факультета социальных наук НИУ ВШЭ, академический руководитель программы «Прикладная политология», e-mail: aakhremenko@hse.ru; **Мячин Алексей Леонидович**, кандидат технических наук, доцент Департамента математики факультета экономических наук Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ, Москва, Россия), старший научный сотрудник Международной научно-учебной лаборатории анализа и выбора решений НИУ ВШЭ, старший научный сотрудник Лаборатории теории выбора и анализа решений Института проблем управления им. В.А. Трапезникова РАН (Москва, Россия), e-mail: amyachin@hse.ru

¹ Исследование выполнено за счет гранта Российского научного фонда (проект № 17-18-01651), Национальный исследовательский университет «Высшая школа экономики».

тературе (насколько известно авторам) описаны особенности паттерн-анализа как метода исследования многомерных массивов данных. Если кластеризация давно и активно используется в политологии, то паттерн-анализ пока еще практически не «встал на вооружение» в нашей науке. При этом паттерн-анализ обладает некоторыми важными и во многом уникальными возможностями.

Было показано, что совместное использование кластерного и паттерн-анализа позволяет выявить согласованные структуры, имеющие ясную интерпретацию в терминах политической науки. Таким образом, в ходе исследования выявлено несколько типов государственной состоятельности, хотя эта задача носила скорее иллюстративный характер.

Эмпирическими индикаторами государственной состоятельности стали доля военных расходов в ВВП, доля военного персонала в общей численности населения, доля налоговых поступлений в ВВП, суммарный уровень убийств и жертв внутренних конфликтов, качество государственных институтов. Данные по более чем 150 странам взяты за 1996, 2005 и 2015 гг. Устойчивые сочетания значений этих показателей, выявленные одновременно с помощью кластерного и паттерн-анализа, формируют искомые структуры государственной состоятельности.

В заключение приводятся наиболее перспективные направления развития описанной методологии. Одним из наиболее важных представляется анализ динамики стран в рамках паттерн-кластерных структур государственной состоятельности.

Ключевые слова: государственная состоятельность; кластер; кластерный анализ; паттерн; анализ паттернов; порядково-инвариантная паттерн-кластеризация.

Для цитирования: Ахременко А.С., Мячин А.Л. Паттерн-анализ и кластеризация в исследовании государственной состоятельности: «Адаптивная оптика» для политической науки // Политическая наука. – М., 2019. – № 3. – С. 112–139. – DOI: <http://www.doi.org/10.31249/poln/2019.03.05>

Введение

Со времен Древней Греции в астрономии известен феномен параллакса – изменения видимого положения небесного тела, наблюдаемого из двух разных пространственных положений. Угол между лучами зрения вкупе с расстоянием между позициями наблюдателя дает возможность с неплохой точностью оценить расстояние до рассматриваемого объекта. В общественных науках «эффект параллакса» – увеличение надежности измерений за счет «переключения» между альтернативными, но родственными аналитическими техниками, – приобретает особую значимость в тех случаях, когда объект исследования искажен противоречивостью концептуальных интерпретаций, неоднозначностью операциональных определений, «шумами» в данных. В этой работе мы по-

кажем, как этот эффект проявляет себя в эмпирическом исследовании государственной состоятельности.

Даже на общем фоне политической науки, с ее обилием сложных для операционализации и измерения понятий, концепт государственной состоятельности выделяется как один из наиболее неоднозначных и «неудобных» для эмпирического анализа [Hendrix, 2010]. На концептуальном уровне государственная состоятельность может трактоваться через призму ключевого доминирующего свойства (например, эффективного государственного аппарата [Charron, 2016]) либо восприниматься как многомерная совокупность неоднородных и, вообще говоря, «равноправных» признаков. В качестве таковых выделяют и качество политических институтов [Hendrix, 2010], и способность собирать налоги (tax extraction) [Hanson, Sigman, 2013], и гарантию прав собственности, и инфраструктурную «проницаемость», обеспечивающую выполнение государственных решений на всей территории страны [Savoia, Sen, 2012], и символическое доминирование для обеспечения национального консенсуса [Wang, 1995]. Каждая из многочисленных концептуальных интерпретаций порождает альтернативные стратегии операционализации – выбора переменных, призванных в комплексе отразить важные для исследователя свойства изучаемого объекта. В случае с государственной состоятельностью и на этом этапе, увы, мы сталкиваемся с существенными проблемами; например, до сих пор нет признанного в сообществе индикатора «экстрактивной» способности государства, и качество институтов почти невозможно измерить без привлечения экспертных – и всегда субъективных – оценок.

Детальный анализ концепта государственной состоятельности не входит в задачи этой работы (развернутый обзор по теме см., например, [Ахременко, Горельский, Мельвиль, 2019 а]). Более того, здесь этот феномен рассматривается в гораздо большей степени как характерная точка приложения усилий в рамках политического исследования, нежели как сущность *per se*. Мы сосредоточимся на проблемах методологического и методического толка и попробуем предложить конкретную исследовательскую стратегию в духе «эффекта параллакса», обеспечения точности фокусировки на устойчивых свойствах сложных политических явлений.

В нашей работе государственная состоятельность будет находиться на «лучах зрения» двух аналитических техник: кластер-

ного и паттерн-анализа. Оба метода нацелены на выявление устойчивых структур в больших массивах неоднородных данных. Решая почти идентичную задачу, они не «дополняют» друг друга в том смысле, в каком количественные и качественные методы могут быть комплементарны в рамках смешанного (mixed) исследовательского дизайна. Напротив, в каком-то смысле это «конкурирующие» инструменты, близкие по цели и замыслу, но довольно существенно отличающиеся «настройкой оптики». Кластерный анализ в большей мере ориентирован на выделение пространственно (с точки зрения расстояния) близких объектов, тогда как паттерн-анализ ищет близость в общих соотношениях параметров данных. Это весьма существенное отличие, которое будет пояснено нами в деталях ниже, в общем случае предполагает значительное несходство в результатах кластерного и паттерн-анализа одной и той же совокупности данных. Однако если обнаруженные этими методами структуры обладают общностью, можно с высокой степенью уверенности утверждать, что в изучаемой реальности мы нашли некоторую существенную закономерность. Так, в данной работе мы покажем, что кластерный анализ и паттерн-анализ уверенно идентифицируют сходные группы стран, однородных по структуре государственной состоятельности.

Исходные данные и выбор методологии

Как было отмечено выше, феномен государственной состоятельности пока не имеет общепринятого (мягко говоря) параметрического описания, и, соответственно, существует неопределенность в выборе базовой системы показателей. Формируя таковую, мы учитывали комплекс различных критериев.

Во-первых, требовалось определить, какую совокупность наблюдений выбрать. Очевидными объектами анализа выступают страны; но интересна также и возможность построения динамических траекторий развития отдельных государств и их групп. Таким образом, в качестве наблюдений выбран тип «страна-год».

Во-вторых, необходимо определиться с выбором параметров и шкал измерения. Мы учитывали такие факторы, как:

– относительно невысокая корреляция между отдельными показателями, характеризующими государственную состоятель-

ность. Этот критерий обусловлен необходимостью избежать дублирования информации в различных измерениях;

- наличие данных за выбранный период времени;
- повышение уровня государственной состоятельности с увеличением отдельно взятого показателя. Поскольку каждый из выбранных индикаторов описан на параметрическом уровне, использовались только количественные шкалы измерения.

В-третьих, существенен вопрос о выборе метрики: каким образом измерять близость между объектами? Мы выбрали метрику Хемминга для паттерн-анализа и евклидово расстояние (простое и квадратичное) для кластеризации. Внешне сугубо «технический», этот вопрос весьма непросто, и ниже мы приведем некоторые соображения по обоснованию такого выбора.

Наконец, существенен вопрос о числе групп, которое должно получиться при разбиении. Поскольку данный вопрос также является открытым, в работе предполагается эндогенное определение как количества групп, так и их состава.

В качестве исходных данных исследованы показатели 166 стран в период 1996–2015 гг. (в 1996 г. – 150; в 2005 – 166; в 2015 г. – 166). В общем виде объекты типа «страна-год» обозначены через s_i . Таким образом, исследуются 482 объекта $s_i \in S$, каждый из которых описывается вектором $s_i = (s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5})$, где:

- s_{i1} – доля военных расходов в ВВП (*Mil_exp*);
- s_{i2} – доля военного персонала в общей численности населения (*Mil_pers*);
- s_{i3} – доля налоговых поступлений в ВВП (*Taxes*);
- s_{i4} – показатель, рассчитываемый как величина, обратная суммарному уровню убийств и жертв внутренних конфликтов (*Safety*);
- s_{i5} – показатель, характеризующий качество государственных институтов (*WGI*). Рассчитывается как первая главная компонента четырех показателей широко известного проекта Всемирного банка World Governance Indicators: контроль над коррупцией, власть закона, эффективность правительства, качество регулирования.

Источники информации по выбранным показателям приводятся в табл. 1.

Таблица 1

**Источники данных по показателям
государственной состоятельности**

<i>Показатель</i>	<i>Источник</i>
Доля военных расходов в ВВП (<i>Mil_exp</i>)	Стокгольмский институт исследования проблем мира (SIPRI)
Доля военного персонала в общей численности населения (<i>Mil_pers</i>)	Международный институт стратегических исследований (IISS)
Показатель, рассчитываемый как величина, обратная суммарному уровню убийств и жертв конфликтов (<i>Safety</i>)	1) Управление ООН по наркотикам и преступности (UNODC); 2) Институт исследований мира в Осло (PRIO)
Доля налоговых поступлений в ВВП (<i>Taxes</i>)	Мировой институт исследований экономического развития (UNU-WIDER)
Качество государственных институтов (<i>WGI</i>)	Всемирный банк (WB).

На следующем после операционализации этапе встает не менее важный вопрос: какую методологию поиска схожих (по выбранной системе показателей) групп стран выбрать? Нами были рассмотрены пять основных подходов, использующихся в такого рода исследованиях:

- составление единого агрегированного рейтинга;
- использование теории индивидуального и коллективного выбора;
- использование методов классификации;
- использование методов кластеризации;
- использование методов анализа паттернов.

Рассмотрим подробнее каждый из них. Составление единого рейтинга стран на базе выбранной системы показателей возможно с использованием некоторой агрегированной оценки. Наиболее популярны средневзвешенная и среднеарифметическая. Первая применяется в случае, когда мы точно можем определить весовой коэффициент каждого показателя, вторая – когда хотим продемонстрировать, что все показатели равнозначны. В таком случае встанут весьма важные вопросы. Как определить вклад каждого отдельно взятого показателя в единый рейтинг? Можем ли мы сказать, какой показатель наиболее важен? Высокие значения одного показателя могут компенсировать низкие значения другого? Можно ли сказать, что все показатели равнозначны? И если нет, как определять весовые коэффициенты? Ответов на эти вопросы нет не только у нас, но и, судя по литературе, ни у кого из исследователей государственной состоятельности. Кроме того, в нашем исследовании

не ставилась задача определения «лучшей» и «худшей» страны в некоторой единой оценке *state capacity*.

Использование теории индивидуального и коллективного выбора имеет свои проблемы. Возможно использование ряда правил, к примеру, Борда, Кондорсе, Нансона, порогового агрегирования. Данный подход позволяет составлять четкое ранжирование стран и имеет ряд преимуществ, однако затрудняет процесс объединения стран в группы, обладающие схожими внутренними структурами. Кроме того, все методы агрегирования дают порядковые, а не интервальные оценки. Другими словами, мы переходим от количественных шкал к номерам наблюдений в ранжированных рядах; на данном этапе трудно оценить, насколько значительны потери информации при таком переходе.

Еще одна опция – использование современных методов классификации, таких как SVM, KNN, «случайный лес» и т.д. Методы классификации данных позволяют проводить разбиение исходных объектов на некоторые классы, количество которых, как правило, должно быть известно заранее. Также необходимо иметь информацию о типичных представителях каждого класса, определить обучающую и тестовую выборки. Данные требования существенно затрудняют использование методов классификации в нашей задаче.

Использование методов кластерного анализа видится перспективным по той причине, что этот метод по своему дизайну исходно ориентирован на поиск структур в многомерных данных. Поскольку конечное количество кластеров мы заранее не знаем (хотя и можно сделать предварительные предположения), а также пытаемся избежать необходимости экспертных оценок о составе каждого кластера перед применением соответствующих методов, использование кластерного анализа целесообразно при решении настоящей задачи. Мы охарактеризуем эту стратегию в деталях ниже.

Как и кластерный анализ, совокупность методов анализа паттернов подходит для выявления устойчивых структур в данных. Целесообразным представляется изучение соотношений между всеми показателями и использование различных мер близости при определении конечного разбиения. Поэтому результаты кластерного анализа предлагается сопоставить с результатами методов анализа паттернов, основанных на парном сравнении показателей. Эта техника будет показана нами в отдельном разделе.

Кластерный анализ

В связи с накоплением больших объемов информации в настоящее время востребована задача поиска закономерностей в разнородных данных. Одним из возможных подходов к разбиению исходного множества объектов на подмножества, содержащие схожие по некоторой мере близости объекты, является кластерный анализ. Само понятие «кластер» в общем случае определяется так [Миркин, 2011, с. 4]: «Под кластером обычно понимается часть данных (в типичном случае – подмножество объектов или подмножество переменных, или подмножество объектов, характеризуемых подмножеством переменных), которая выделяется из остальной части наличием некоторой однородности элементов». Методам кластерного анализа посвящен ряд работ (в частности, [Jain, Murty, Flynn, 1999; Xu, Wunsch, 2005]), в которых описываются типовые метрики, классификация методов кластерного анализа, а также даны рекомендации по их использованию в конкретных случаях. Для кластера можно выделить несколько характеристик, среди которых в настоящей работе ключевыми являются внешняя изолированность и внутренняя однородность. Другими словами, расстояния внутри кластеров должны быть минимальны, а расстояния между кластерами – максимальны.

Одна из главных целей применения методов кластерного анализа – улучшение понимания структуры объектов, описанных многомерными данными. При этом может отсутствовать исходная гипотеза о количестве и составе кластеров, хотя могут быть выдвинуты некоторые предположения, исходя из экспертных оценок и применения стандартных методов теории вероятностей и математической статистики. Именно в связи с тем, что точное количество (как и состав) групп стран заранее неизвестно, в работе используется кластерный анализ.

Существует множество способов классифицировать алгоритмы кластеризации. К примеру, часто выделяют следующие группы¹:

- алгоритмы, основанные на разделении данных (включая итеративные);
- иерархические алгоритмы;

¹ Данная классификация не является единственно возможной.

- модельные алгоритмы;
- алгоритмы, использующие концентрацию объектов;
- алгоритмы, основанные на квантовании объектов в грид-структуры.

При всем разнообразии подходов нужно ответить на вопрос: какой метод выбрать? И имеется ли лучший метод кластеризации? Разумеется, данным вопросом задавались исследователи во многих работах. Приведем несколько уже ставших классическими примеров. В [Dubes, Jain, 1976] дан набор их 15 точек, представленных на рис. 1.

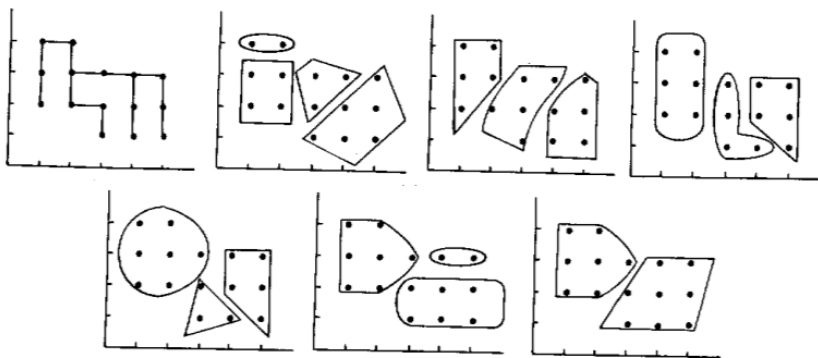


Рис. 1

**Результаты применения различных методов
к одному набору данных**

*Пример из [Dubes, Jain 1976].

На рис. 1 показан классический пример влияния выбора метода кластеризации на конечный результат. Однако набор данных весьма небольшой – 15 точек. Возможно ли, что на больших выборках такой эффект не наблюдается?

Чтобы ответить на данный вопрос, воспользуемся библиотекой `scikit-learn` для языка Python. В данной библиотеке представлены различные методы машинного обучения и встроены классические наборы данных. Описание к библиотеке имеется в документации¹.

¹ Comparing different clustering algorithms on toy datasets // `scikit-learn`. – Режим доступа: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_

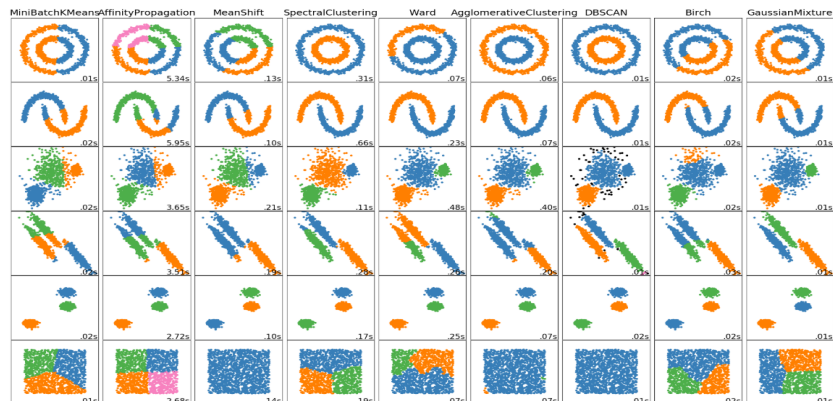


Рис. 2.

Результаты разбиения при использовании различных методов кластеризации

Рис. 1–2 демонстрируют, что использование различных методов может приводить к различным результатам. Поэтому возникает вопрос выбора алгоритма для конкретной задачи. Ответ на него требует привлечения дополнительной информации, исходя из специфики рассматриваемой проблемы. В настоящей работе предложен следующий подход.

– Взяв за основу иерархическую кластеризацию, обеспечить максимальную устойчивость разбиения за счет использования широкого набора комбинаций метрик (евклидово расстояние, квадрат евклидова расстояния) и правил объединения (межгрупповая связь, внутригрупповая связь, дальние соседи, метод Варда, метод центроидов). Принадлежащими одной типологической группе мы будем считать только страны, одинаково кластеризованные всеми используемыми алгоритмами. Детально такой подход раскрыт в работе [Ахременко, Горельский, Мельвиль, 2019 б].

– Комбинирование кластерного анализа с методами анализа паттернов, основанными на парном сравнении показателей с независимыми от выбора исходной последовательности данных.

Анализ паттернов

Анализ паттернов – современный метод выявления закономерностей среди больших массивов разнородных данных; он гораздо хуже, по сравнению с кластерным анализом, известен в политической науке, и мы охарактеризуем его более детально. В [Анализ паттернов... 2013, с. 3] приведено следующее определение: «Анализ паттернов – это новая область анализа данных, связанная с поиском взаимосвязей исследуемых объектов, построением их классификации и исследованием развития объектов во времени». Однако термин «паттерн» определяется по-разному – в зависимости от области применения. К примеру, в [Shawe-Taylor, Cristianini, 2004] под анализом паттернов понимается «процесс нахождения общих соотношений в наборе данных», а под «паттерном» – «любые отношения, закономерности или структура, присущая некоторому набору данных». В [Анализ паттернов... 2013, с. 4] под «паттерном» предложено понимать «комбинацию определенных, с точностью до погрешности, значений некоторого подмножества признаков, что объекты с этими значениями достаточно сильно отличаются от других объектов». В [Анализ данных науки... 2012, с. 6] под паттерном понимается «набор значений системы показателей, описывающих какую-либо группу объектов, а также саму группу объектов, имеющих такие же или почти такие же значения показателей данной системы». Для определенности в работе будем использовать определение из [Мячин, 2019, с. 139]: «комбинация определенных качественно похожих признаков». Кратко приведем формальное описание метода.

Как описано выше, исследуется множество объектов типа «страна-год» $s_i \in S$: $|S|=482$, каждому из которых поставлен во взаимно однозначное соответствие вектор $s_i = (s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5})$. Задача состоит в том, чтобы, используя определенную меру близости $p(s_i, s_k)$, объединить структурно схожие объекты в единую группу. Визуализируются результаты с использованием системы параллельных координат, которая состоит из равномерно распределенных (как правило, вертикальных) осей, каждая из которых характеризует один из исследуемых показателей. Поскольку у каждого изучаемого объекта задано признаковое описание, возможно построить некоторую кривую, проходящую через $s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5}$ (другими словами, функцию $\varphi^s: R \rightarrow R$, причем $\varphi^s(j) = s_{ij} \forall j = 1, \dots, 5$).

В настоящей работе используется кусочно-линейная функция, в связи с чем $\varphi^s(\alpha) = \{\alpha q_j^s + \beta_j^s\}$ при $j \leq \alpha \leq j+1$; $q_j^s j + \beta_j^s = s_{ij}$; $q_j^s(j+1) + \beta_j^s = s_{ij+1} \forall j = 1, 2, 3, 4$.

В качестве меры близости возможно использование различных метрик, к примеру, описанных в [Aleskerov, Ersel, Yolalan, 1997]. В настоящей работе используются методы анализа паттернов, основанные на парном сравнении показателей [Мячин, 2016; Мячин, 2019 а], исходя из следующих особенностей исходных данных:

- в связи с отсутствием четкого описания набора данных, характеризующих государственную состоятельность отдельных стран, и использованием системы параллельных координат для визуализации данных требуются методы анализа паттернов, результаты которых не будут никоим образом зависеть (в том числе и при корректной визуализации данных) от выбора последовательности выходных параметров;

- ввиду отсутствия исходного предположения о необходимом количестве групп в итоговом разбиении предполагается эндогенное определение их количества (как и состава);

- сравнение результатов различных методов предполагает относительно невысокую вычислительную сложность, позволяющую в ограниченные сроки произвести необходимые расчеты и сопоставить результаты.

Приведем пример использования методов анализа паттернов (в общем виде). Пусть известны гипотетические данные по пяти параметрам, взятые по пяти странам (см. табл. 2). Предполагается, что данные предварительно проанализированы: исследованы выбросы, проведен корреляционный анализ, пропуски либо заполнены, либо удалены.

Таблица 2

Пример гипотетических данных по пяти странам (за один год)

Страна	<i>Mil exp</i>	<i>Mil pers</i>	<i>Taxes</i>	<i>Safety</i>	<i>WGI</i>
Страна 1	0,45	0,15	0,45	0,15	0,45
Страна 2	0,2	0,55	0,2	0,5	0,2
Страна 3	0,5	0,17	0,5	0,21	0,47
Страна 4	0,55	0,2	0,53	0,23	0,52
Страна 5	0,17	0,6	0,17	0,55	0,14

Приведем визуализацию данных с использованием двух подходов: системы параллельных координат и многоугольника оценки (рис. 3). Отметим, что рассматриваемые подходы в данном

случае – не более чем методы наглядного представления каждой из исследуемых стран. При использовании методов анализа паттернов классическим представлением являются кусочно-линейные функции в системе параллельных координат. Данное представление позволяет наглядно выявлять объекты (в нашем случае страны), имеющие схожие структуры данных, количественные показатели которых могут существенным образом различаться.

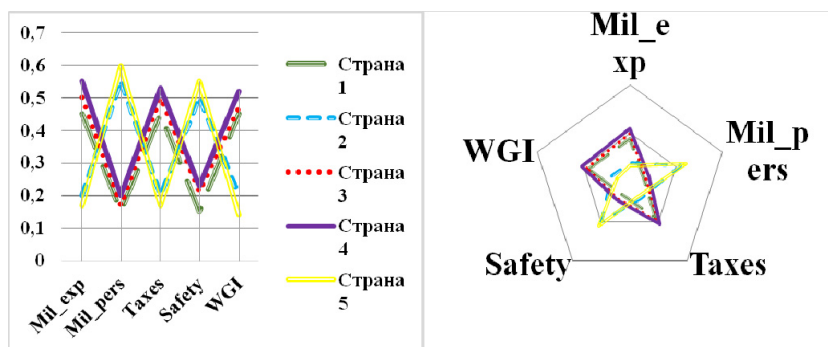


Рис. 3.

**Визуализация гипотетических данных
(по данным из табл. 2) с использованием системы
параллельных координат (слева)
и многоугольника оценки (справа)**

На рис. 3 наглядно показано, что весьма четко выделяются две группы стран: (страна 1, страна 3, страна 4) и (страна 2, страна 5). Данные группы схожи не только по структурам входящих в них стран, но и по количественным значениям всех изучаемых показателей. Методы анализа паттернов должны разбить данные страны строго на две группы.

В таком случае логичными кажутся следующие вопросы:

- 1) В чем заключается отличие от классических методов кластерного анализа?
- 2) Каким образом можно автоматически провести данное разбиение?

Дадим краткие ответы на оба вопроса. Несмотря на действительную схожесть с классическими методами кластерного

анализа, имеется принципиальное отличие не только в используемых метриках, но и в конечных результатах. Методы кластерного анализа, как правило, объединяют в единый кластер объекты с близкими значениями отдельных показателей. Для методов анализа паттернов принципиален поиск схожих структур данных (соотношений значений признаков). К примеру, возьмем два гипотетических объекта $s_a = (20, 10, 30)$ и $s_b = (200, 100, 300)$. Очевидно, что показатели данных объектов существенным образом различаются. Однако их структура весьма схожа: показатели второго объекта есть показатели первого, помноженные на 10, и, соответственно, анализ паттернов выявит эту схожесть структур (несмотря на разницу абсолютных значений показателей).

Приведенный пример демонстрирует, что использование таких традиционных метрик, как евклидово расстояние, в данном случае будет разбивать объекты на две группы. В связи с этим приведем (очень кратко) один из возможных подходов, выбранных в данной работе, – порядково-инвариантную паттерн-кластеризацию, описанную в: [Мячин, 2019]. Первым шагом формируется матрица парных сравнений X_i

$$X_i = \begin{pmatrix} x_{11}^i & x_{12}^i & \dots & x_{15}^i \\ x_{21}^i & x_{22}^i & \dots & x_{25}^i \\ \dots & \dots & \dots & \dots \\ x_{51}^i & x_{m2}^i & \dots & x_{55}^i \end{pmatrix},$$

где каждый элемент матрицы определяется как

$$x_{vj}^i = \begin{cases} 1, \text{если } _s_{iv} < s_{ij}, \\ 0, \text{если } _s_{iv} = s_{ij}, \\ 2, \text{если } _s_{iv} > s_{ij}. \end{cases}$$

При работе с числовыми шкалами (исходя из выбранных для исследования показателей) любые парные сравнения будут опре-

делены однозначно, в связи с чем могут быть использованы только элементы, находящиеся выше (или ниже) главной диагонали матрицы. Далее формируется десятичный позиционный код на основе матриц парных сравнений и рассчитывается расстояние Хемминга между полученными кодировками. При нулевом расстоянии Хемминга объекты объединяются в единую группу, в любом другом случае – разделяются (подробнее см.: [Мячин, 2019 а]). У данного подхода есть одно существенное преимущество, определившее его использование в данной работе, – независимость конечного разбиения от последовательности входных данных.

Однако описанный выше подход не лишен существенного недостатка: как корректировать результаты при наличии погрешности в данных? Очевидно, что при парном сравнении показателей наличие определенных неточностей может повлиять на результат. Поэтому с данными проводится предварительная работа, к примеру, оцениваются выбросы. Несмотря на это, желательно иметь возможность автоматической корректировки результатов для упрощения процесса интерпретации конечного разбиения. Одним из возможных подходов (используемых в данной работе), является вычисление центроидов каждой группы для минимизации возможности ошибки. Центроид каждой группы вычисляется как

$$s_{average}^g = \frac{1}{|g|} \sum_{i=1}^{|g|} s_i,$$

где g – конкретная группа, полученная при использовании порядково-инвариантной паттерн-кластеризации;

$|g|$ – количество объектов, входящих в данную группу.

Отметим, что в работе [Мячин, 2019 b] показано, что $s_{average}^g$ будет принадлежать к той же группе, что и объекты, на основе значений которых он образован. Далее вычисляется расстояние всех объектов до каждого центроида согласно формуле

$$p^v(s_i, s_{average}^g) = \sqrt{\sum_{j=1}^5 (p_{ij} - s_{average,j}^g)^2}.$$

Критерием корректировки результатов служит минимизация расстояния $p^v(s_i, s_{average}^g)$, т.е. величина

$$z_i = \min(p^1(s_i, x_{average}^1), \dots, p^h(s_i, s_{average}^h))$$

и, при $p^v(s_i, s_{average}^g) = z_i$, объект относится к той же группе, к которой принадлежит соответствующий центроид.

Результаты

Приведем некоторые из полученных в работе результатов. В большинстве случаев методы анализа паттернов и кластерного анализа, несмотря на принципиальную разницу в используемых метриках, привели к схожим и хорошо интерпретируемым результатам, что указывает на корректность полученного разбиения. Отметим, что в данной работе не ставилась цель построения обобщенного рейтинга стран по государственной состоятельности и приведения характеристик «лучше / хуже». Приведенные ниже результаты демонстрируют схожесть стран по совокупности исследуемых показателей, характеризующих государственную состоятельность.

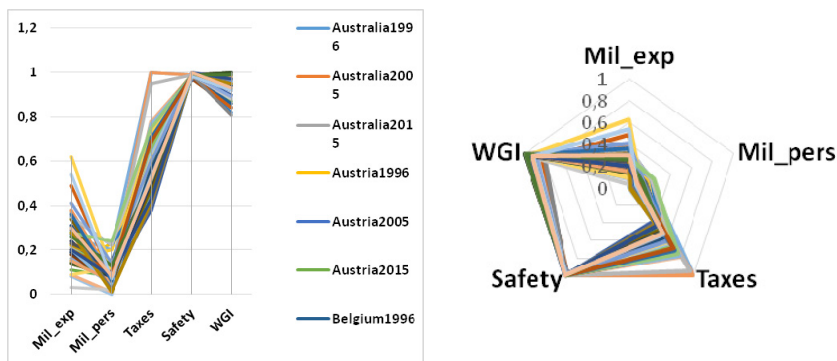


Рис. 4.
Визуализация группы объектов «А»

Группа «А» представляет собой «западную классику», представленную 50 объектами, описывающими 17 стран: Австрию, Австралию, Данию, Финляндию, Бельгию, Канаду, Исландию, Ирландию, Люксембург, Швецию, Швейцарию, Великобританию, Нидерланды, Францию, Германию, Новую Зеландию и Норвегию. Для них характерны высокие значения показателей доли налоговых поступлений и качества государственных институтов, низкие значения уровня убийств и жертв конфликтов², а также низкие и относительно невысокие значения показателей, характеризующих долю военного персонала в общей численности населения и долю военных расходов в ВВП страны. Все страны данной группы не меняют принадлежности к исследуемому паттерну / кластеру за весь исследуемый период, за исключением Ирландии в 2015 г.

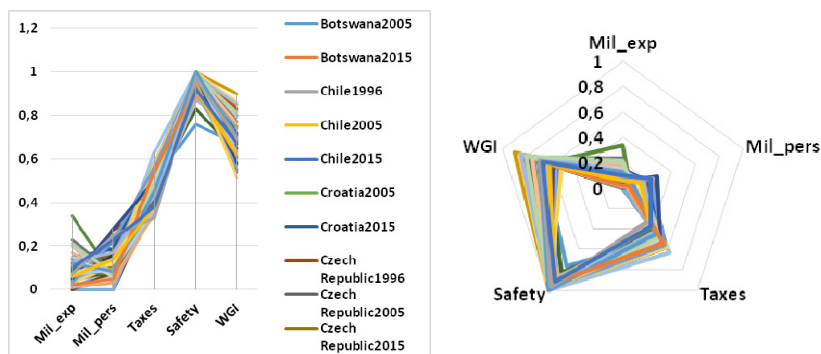


Рис. 5.
Визуализация группы «В»

Группа «В» имеет ряд общих черт с предыдущей группой, к примеру, высокие показатели безопасности и низкие значения доли военных расходов. Однако наблюдаются меньшие (в сравнении с группой «А») значения качества государственных институтов и доли налоговых поступлений. Такая структура характерна для 24 стран: Ботсвана (2005–2015), Чили (1996–2015), Чехия (1996–2015), Хорватия (2005–2015), Доминиканская

² Как отмечено ранее, показатель Safety является величиной, обратной суммарному уровню убийств и жертв конфликтов.

Республика (1996–2015), Эстония (1996–2015), Фиджи (1996), Грузия (2015), Венгрия (1996–2015), Ирландия (2015), Италия (1996–2015), Латвия (1996–2015), Литва (1996, 2015), Малайзия (1996), Мальта (1996–2015), Маврикий (2005–2015), Черногория (2015), Польша (1996–2015), Португалия (1996–2015), Словакия (1996–2015), Словения (1996–2015), Испания (1996–2015), Тринидад и Тобаго (1996), Уругвай (1996–2015). Эту группу можно охарактеризовать как «страны догоняющего развития».

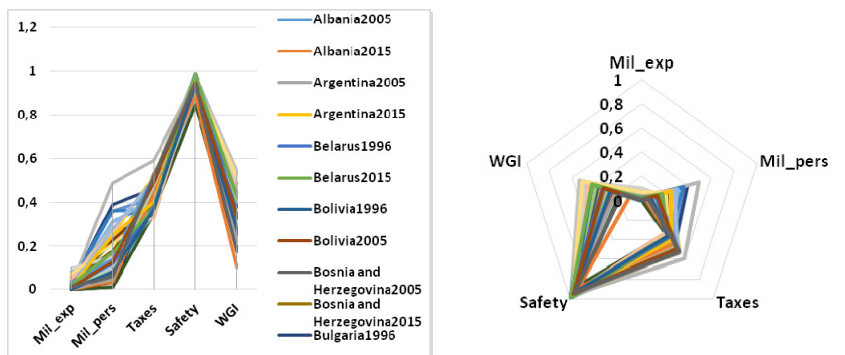


Рис. 6.
Визуализация группы «С»

Группа «С» представлена 63 объектами, характерными для 34 стран. Среди них Албания (2005, 2015), Аргентина (2005, 2015), Беларусь (1996, 2015), Боливия (1996, 2005), Босния и Герцеговина (2005–2015), Болгария (2005–2015), Египет (1996), Эритрея (1996), Фиджи (2005–2015), Габон (1996), Гвинея (2015), Казахстан (2015), Киргизия (2005–2015), Македония (1996–2005), Джибути (1996–2015), Мозамбик (2015), Румыния (1996–2015), Сербия (2005–2015), Суринам (1996–2005), Таджикистан (2015), Турция (1996–2015), Украина (1996–2015), Узбекистан (2005–2015), Вьетнам (2005–2015), Замбия (1996), Зимбабве (1996, 2015), Таиланд (2015), Того (2015) и Тунис (1996–2015). Данную группу отличают более низкие (по сравнению с группами «А» и «В») показатели качества государственных институтов, относительно невысокие доли военных расходов в ВВП страны.

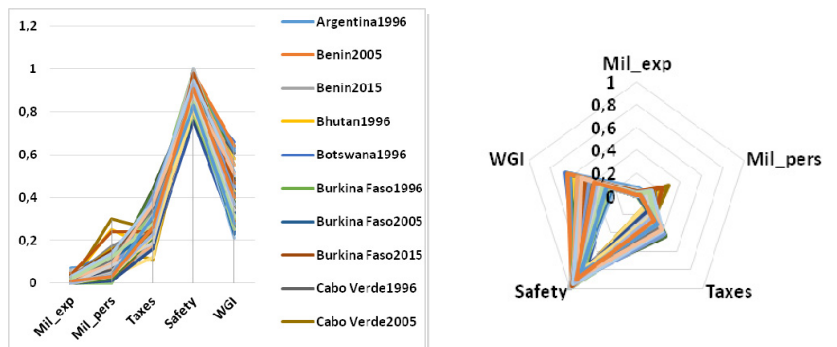


Рис. 7.
Визуализация группы «D»

Данная группа весьма сходна с предыдущей, однако можно отметить отличия в соотношениях показателей, характеризующих собираемость налогов и удельную долю военнослужащих, а также в качестве государственных институтов. При необходимости уменьшения количества групп (если того требует постановка задачи) возможно объединение групп «C» и «D», однако в рамках данного исследования и сопоставления результатов паттерн-анализа и кластеризации предлагается сохранить описанное разбиение. Группа «D» состоит из 110 объектов, описывающих такие страны, как Аргентина (1996), Бенин (2005–2015), Бутан (1996), Ботсвана (1996), Буркина-Фасо (1996–2015), Кабо-Верде (1996–2015), Камерун (2015), Коста-Рика (1996–2015), Кот-д'Ивуар, Доминиканская Республика (1996, 2015), Эквадор (1996–2015), Эфиопия (2005–2015), Габон (2005–2015), Гамбия (1995–2015), Грузия (2005), Гана (1996–2015), Гайана (1996), Индонезия (1996–2015), Казахстан (1996), Кения (1996–2015).

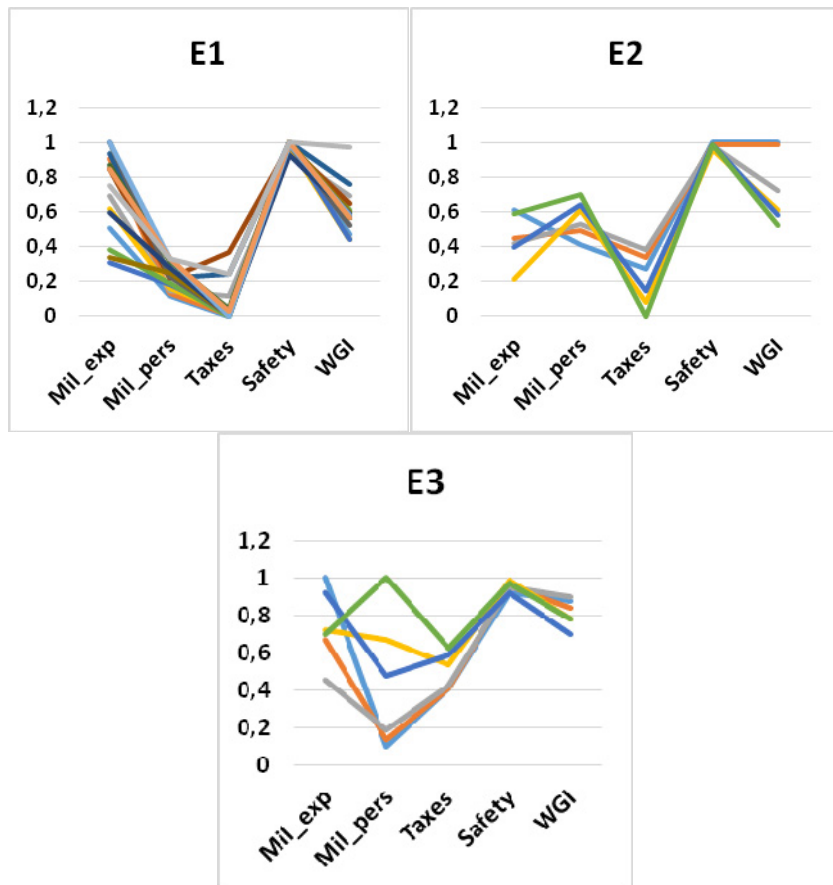


Рис. 8

Визуализация группы «Е».
Пример получения различных результатов

Группа «Е» служит хорошим примером получения хоть и различных, но согласованных результатов. С использованием методов кластерного анализа данные объекты объединяются в единый кластер, в то время как с использованием методов анализа паттернов разбиваются на три подгруппы. Объяснение данному факту весьма простое: выбранные для работы методы кластерного

анализа в качестве меры близости используют евклидово расстояние, методы анализа паттернов – расстояние Хемминга между кодировками, полученными при парном сравнении показателей с последующей корректировкой результатов при помощи вычисления центроида группы (что важно для группы «Е3»).

В подгруппе «Е1» результаты хорошо интерпретируются: относительно высокие показатели доли военных расходов, средние – безопасности и качества государственных институтов. В данную подгруппу вошли почти исключительно «нефтяные монархии» Ближнего Востока: Бахрейн (2005–2015), Кувейт (1996–2015), Оман (2005–2015), Катар (2005–2015), Саудовская Аравия (1996–2015), ОАЭ (2005–2015). В подгруппе «Е2» можно наблюдать увеличение значений показателя, характеризующего доли военного персонала. Данная подгруппа характерна для почти того же набора стран, но в другие периоды времени: Бахрейн (1996), Бруней (1996), Оман (1996), Катар (1996). В подгруппе «Е3» видно значительное увеличение доли налоговых поступлений в ВВП страны, что характерно для США (1996–2015) и Израиля (1996–2015).

Группа «F» – другой хороший пример получения немного отличных, но весьма согласованных результатов при использовании выбранных для исследования методов анализа паттернов и кластерного анализа. Методы анализа паттернов, в связи со спецификой выбранной меры близости, выделяют три различных подгруппы, в то время как методы кластерного анализа объединяют все объекты в единую группу. Для подгруппы «F1» характерны относительно невысокие налоговые поступления, средние и относительно низкие показатели качества государственных институтов, крайне низкий показатель безопасности. В данную подгруппу включены Сальвадор (1996–2015) и Ямайка (2005). В подгруппе «F2» можно отметить более высокие (в сравнении с «F1», но низкие – в сравнении со многими другими странами) показатели безопасности, а также низкие и крайне низкие показатели качества государственных институтов. В «F2» вошли Ирак (2005), Демократическая Республика Конго (1996) и Афганистан (2015). Подгруппу «F3» отличают более высокие налоговые поступления. К ней относятся Таджикистан (1996), Гватемала (2005), Бурунди (1996), Колумбия (1996), Шри-Ланка (1996), Гондурас (2005–2015), Венесуэла (2005–2015) и Колумбия (2005). Все эти государства в той или иной мере находятся на грани state failure.

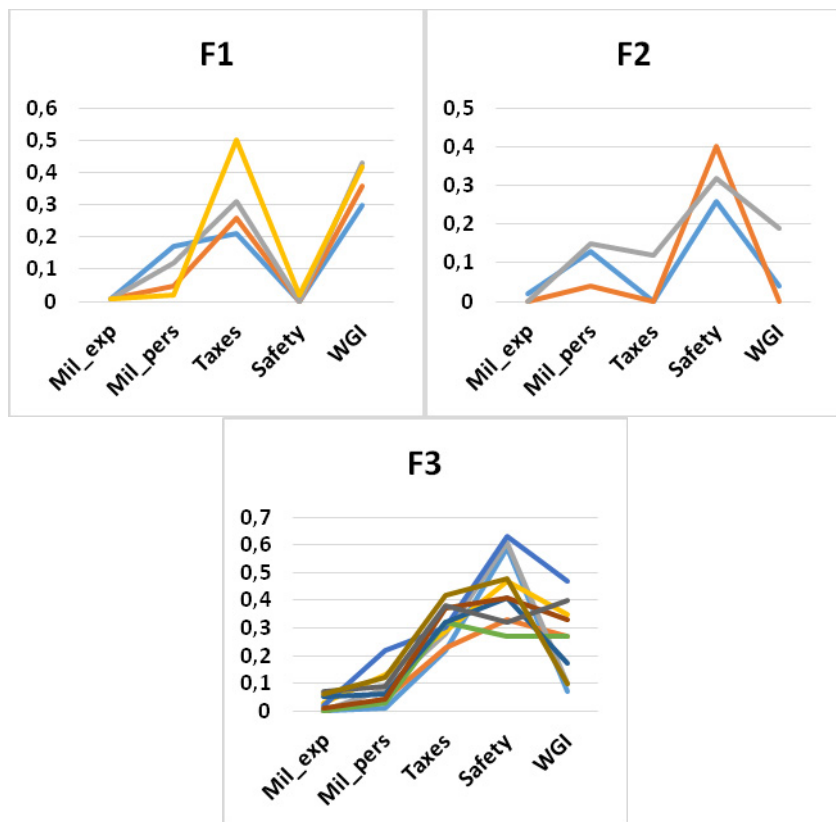


Рис. 9.

Визуализация группы «F».

Пример получения различных результатов

Для сопоставления приведем средние значения по каждой группе на рис. 10.

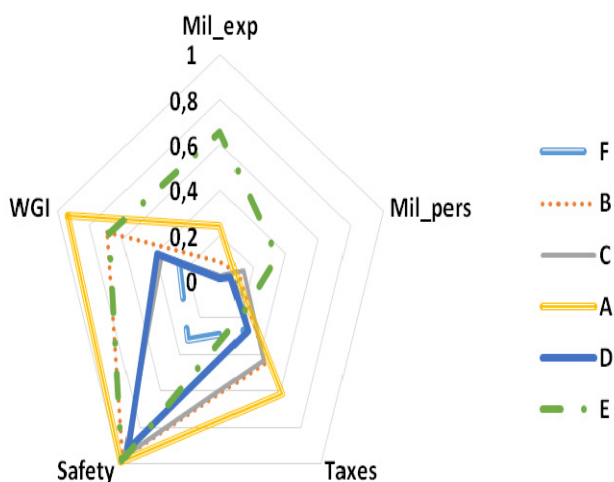


Рис. 10.
Визуализация полученных групп.
Сопоставление результатов

На рисунке 10 видно, что каждая из выделенных групп имеет характерные «очертания» в пространстве пяти рассматриваемых признаков. Полученные наглядные структуры могут, с нашей точки зрения, служить эффективным инструментом сравнительного политологического анализа.

Заключение

Центральный фокус этой работы – методологический. На примере набора индикаторов государственной состоятельности мы постарались показать конкретную стратегию выявления устойчивых структур в многомерных массивах данных. Ключевая особенность этой стратегии – применение родственных, но существенно различающихся по своим техническим особенностям многомерных методов – кластерного и паттерн-анализа. Было показано, что их совместное использование позволяет выявить согласованные

структуры, имеющие ясную интерпретацию в терминах политической науки. Таким образом, в ходе нашего исследования было выявлено несколько типов государственной состоятельности, хотя эта задача носила для нас скорее иллюстративный характер.

Попутно были впервые в политологической литературе (насколько нам известно) показаны особенности паттерн-анализа как метода исследования многомерных массивов данных. Если кластеризация давно и активно используется в политологии, то паттерн-анализ пока еще практически не «встал на вооружение» в нашей науке. При этом паттерн-анализ обладает некоторыми важными и во многом уникальными возможностями.

Делая акцент на анализе структур (устойчивых взаимных соотношений объектов и признаков), мы фактически ставим более широкую методологическую проблему: а в достаточной ли мере используются возможности такого анализа в современной политической науке? На сегодняшний день господствующим трендом развития политической методологии является анализ причинности (не только каузальный анализ в узком смысле, но вся совокупность техник и инструментов выявления причинно-следственных связей). Важность данного аспекта не вызывает сомнений, однако можно ли свести все многообразие эмпирических исследований к решению этой задачи? Тем более что в политике многие процессы де-факто не могут быть однозначно разложены на «причины» и «следствия», так как многие из них содержат в себе петли обратных связей. Например, влияние репрессий на протестную активность с трудом поддается причинному анализу, так как здесь имеется «встроенная эндогенность»: репрессии являются реакцией на протест, который на своем следующем витке «учитывает» репрессивную реакцию властей, и т.д.

Конкретные примеры исследовательских вопросов, прямо вытекающих из нашего исследования и при этом находящиеся вне жестких рамок каузального подхода, можно сформулировать следующим образом. Если некоторые страны обладают «переходным» характером государственной состоятельности (принадлежат к разным структурам в период с 1996 по 2015 г.), а другие демонстрируют стабильные структуры, то с чем такие различия могут быть связаны? Какие структуры могут «переходить» друг в друга, а какие жестко изолированы? Решение последней проблемы имело бы и вполне конкретное прогностическое значение.

Именно эти вопросы в первую очередь очерчивают перспективы дальнейших исследований в рамках выбранной темы. Следующим шагом, таким образом, должен стать анализ динамики стран в рамках паттерн-кластерных структур государственной состоятельности.

Список литературы

- Анализ данных науки, образования и инновационной деятельности с использованием методов анализа паттернов: препринт WP7/2012/07 / Ф.Т. Алескеров, Л.М. Гохберг, Л.Г. Егорова, А.Л. Мячин, Г.С. Сагиева; Нац. исслед. ун-т «Высшая школа экономики». – М.: Изд. дом Высшей школы экономики, 2012. – 72 с. – Режим доступа: https://wp.hse.ru/data/2013/01/22/1305726760/WP7_2012_07_f_.pdf (Дата обращения: 15.05.2019.)
- Анализ паттернов в статике и динамике, часть 1: обзор литературы и уточнение понятия / Ф.Т. Алескеров, В.Ю. Белоусова, Л.Г. Егорова, Б.Г. Миркин // Бизнес-информатика. – М., 2013. – № 3(25). – С. 3–18.
- Ахременко А.С., Горельский И.Е., Мельвиль А.Ю. Как и зачем измерять и сравнивать государственную состоятельность различных стран мира? Теоретико-методологические основания // Полис. Политические исследования. – М., 2019 а. – № 2. – С. 8–23. – DOI: <https://doi.org/10.17976/jpps/2019.02.02>
- Ахременко А.С., Горельский И.Е., Мельвиль А.Ю. Как и зачем измерять и сравнивать государственную состоятельность различных стран мира? Опыт эмпирического исследования // Полис. Политические исследования. – М., 2019 б. – № 3. – С. 49–68. – DOI: <https://doi.org/10.17976/jpps/2019.03.04>
- Миркин Б.Г. Методы кластер-анализа для поддержки принятия решений: обзор: препринт WP7/2011/03. – М.: Изд. дом Национального исследовательского университета «Высшая школа экономики», 2011. – 88 с. – Режим доступа: https://www.hse.ru/data/2011/05/19/1213868030/WP7_2011_03f.pdf (Дата помещения: 15.05.2019.)
- Мячин А.Л. Анализ паттернов в системе параллельных координат на базе парного сравнения показателей // Автоматика и телемеханика. – М., 2019 а. – № 1. – С. 138–152. – DOI: <https://doi.org/10.1134/S0005231019010100>
- Мячин А.Л. Анализ паттернов: диффузионно-инвариантная паттерн-кластеризация // Проблемы управления. – 2016. – № 4. – С. 2–9.
- Мячин А.Л. Определение центроидов для повышения точности порядково-инвариантной паттерн-кластеризации // Управление большими системами. – 2019 б. – № 78. – С. 6–22.
- Aleskerov F., Ersel H., Yolalan R. Clustering Turkish commercial banks according to structural similarities // Yapi Kredi Research Department Discussion Paper Series. – Istanbul, 1997. – P. 97–102.

- Charron N.* Diverging Cohesion? Globalisation, State Capacity and Regional Inequalities Within and Across European Countries // *European Urban and Regional Studies*. – L., 2016. – Vol. 23, N 3. – P. 355–373.
- Dubes R., Jain A.K.* Clustering techniques: the user's dilemma // *Pattern Recognition*. – N.Y., 1976. – T. 8, N 4. – P. 247–260.
- Hanson J., Sigman R.* Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research. – 2013. – Mode of access: http://www-personal.umich.edu/~jkhanson/resources/hanson_sigman13.pdf (accessed: 20.04.2019.)
- Hendrix C.* Measuring State Capacity: Theoretical and Empirical Implications for the Study of Civil Conflict // *Journal of Peace Research*. – L., 2010. – Vol. 47, N. 3 – P. 273–285.
- Jain A.K., Murty M.N., Flynn P.J.* Data clustering: a review // *ACM computing surveys (CSUR)*. – N.Y., 1999. – T. 31, N 3. – P. 264–323.
- Savoia A., Sen K.* Measurement and evolution of state capacity: Exploring a lesser known aspect of governance. – Manchester, UK, 2012. – 28 p. – (ESID Working Paper; N 10).
- Shawe-Taylor J., Cristianini N.* Kernel methods for pattern analysis. – Cambridge: Cambridge univ. press, 2004. – Mode of access: http://read.pudn.com/downloads190/ebook/893343/Kernel_Methods_for_Pattern_Analysis/0521813972.pdf (accessed: 20.04.2019.)
- Wang S.* The Rise of the Regions: Fiscal Reform and the Decline of Central State Capacity in China // *The Waning of the Communist State: Economic Origins of Political Decline in China and Hungary* / Ed. by A. Walder. – Berkeley: Univ. of California press, 1995. – P. 87–114.
- Xu R., Wunsch D.C.* Survey of clustering algorithms // *IEEE Transactions on Neural Networks*. – N.Y., 2005. – Vol. 16, N 3. – P. 645–678. – DOI: <https://doi.org/10.1109/TNN.2005.845141>

Akhremenko A.S., Myachin A.L.*

**Pattern Analysis and Clustering in the Study of State Capacity:
«Adaptive Optics» for Political Science**

The central focus of this paper is a methodological one. Using the set of indicators of state capacity, we demonstrate a specific strategy for identifying sustainable structures in multidimensional data sets that reflect complex and ambiguous concepts of political science. A key feature of this strategy is the application of related, but significantly different technically, multidimensional methods – cluster and pattern analyses. We use hierarchical clustering with various combinations of metrics and amalgamation

* **Akhremenko Andrey**, National Research University Higher School of Economics (Moscow, Russia), e-mail: aakhremenko@hse.ru; **Myachin Alexey**, National Research University Higher School of Economics (Moscow, Russia), e-mail: amyachin@hse.ru

rules, as well as ordinal-invariant pattern-clustering. Properties of pattern analysis as a method for studying multidimensional data are shown for the first time (to the best of our knowledge) in the political science literature. Since clustering has been actively used in political science for a long time, pattern analysis is still practically not adopted in our science. This situation requires correction, since pattern-analysis has some important and in many ways unique capabilities.

It was shown that the combination of pattern and cluster analyses makes it possible to identify consistent structures that have a clear interpretation in terms of political science. Thus, in the course of our study, several types of state capacity were identified (although this task was rather illustrative for us).

We use a set of empirical indicators of state capacity: the share of military spending in GDP, the share of military personnel in the total population, the share of tax revenues in GDP, the total rate of homicides and victims of internal conflicts, and the quality of government institutions. Data for more than 150 countries are taken for 1996, 2005 and 2015. Stable combinations of the values of these indicators, identified simultaneously via pattern and cluster analyses, form the structures of state capacity.

In conclusion, we show the most promising directions for the development of the methodology described in this paper. One of the most important is the analysis of the dynamics of countries within the pattern-cluster structures of state capacity.

Keywords: state capacity; cluster; cluster analysis; pattern; pattern analysis; ordinal-invariant pattern-clustering.

For citation: Akhremenko A.S., Myachin A.L. Pattern Analysis and Clustering in the Study of State Capacity: «Adaptive Optics» for Political Science. *Political science (RU)*. 2019, N 3, P. XX-XX. DOI: <http://www.doi.org/10.31249/poln/2019.03.05>

References

- Akhremenko A.S., Gorelskiy I.E., Melville A. Yu. How and Why Should We Measure and Compare State Capacity of Different Countries? Theoretical and Methodological Foundations. *Polis. Political Studies*. 2019, N 2, P. 8–23. (In Russ.) DOI: <https://doi.org/10.17976/jpps/2019.02.02> (In Russ.)
- Akhremenko A.S., Gorelskiy I.E., Melville A. Yu. How and Why Should We Measure and Compare State Capacity of Different Countries? An Experiment with Empirical Research. *Polis. Political Studies*. 2019, N 3. P. 49–68. DOI: <https://doi.org/10.17976/jpps/2019.03.04> (In Russ.)
- Aleskerov F., Ersel H., Yolalan R. Clustering Turkish commercial banks according to structural similarities. In: *Yapi Kredi Research Department Discussion Paper Series*. Istanbul, 1997, P. 97–102.
- Aleskerov F.T. et al. Data analysis of science, education and innovation using the methods of pattern analysis: preprint WP7/2012/07. Moscow, 2012, 72 p. Mode of access: https://wp.hse.ru/data/2013/01/22/1305726760/WP7_2012_07_f.pdf (accessed: 15.05.2019.) (In Russ.)

- Aleskerov F.T. et al. Methods of pattern analysis in statics and dynamics, part 2: Examples of application for social and economic processes analysis. *Business informatics*. 2013, N 3(25), P. 3–18. (In Russ.)
- Charron N. Diverging Cohesion? Globalisation, State Capacity and Regional Inequalities Within and Across European Countries. *European Urban and Regional Studies*. 2016, Vol. 23, N 3, P. 355–373.
- Dubes R., Jain A.K. Clustering techniques: the user's dilemma. *Pattern Recognition*. 1976, T. 8, N 4. P. 247–260.
- Hanson J., Sigman R. Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research. 2013. Mode of access: http://www-personal.umich.edu/~jkhanson/resources/hanson_sigman13.pdf (accessed: 20.04.2019.)
- Hendrix C. Measuring State Capacity: Theoretical and Empirical Implications for the Study of Civil Conflict. *Journal of Peace Research*. 2010, Vol. 47, N 3, P. 273–285.
- Jain A.K., Murty M.N., Flynn P.J. Data clustering: a review. *ACM computing surveys (CSUR)*. 1999, T. 31, N 3, P. 264–323.
- Mirkin B.G. Cluster analysis methods for decision-making support: overview: preprint WP7/2011/03. Moscow, 2011, 88 p. Mode of access: https://www.hse.ru/data/2011/05/19/1213868030/WP7_2011_03f.pdf (accessed: 15.05.2019.) (In Russ.)
- Myachin A.L. Analysis of patterns in the system of parallel coordinates based on pairwise comparison of indicators. *Automation and Remote Control*. 2019, N 1, P. 138–152. (In Russ.) <https://doi.org/10.1134/S0005231019010100>
- Myachin A.L. Determination of centroids to increase the accuracy of ordinal-invariant pattern clustering. *Upravleniye bol'shimi sistemami*. 2019, N 78, P. 6–22. (In Russ.)
- Myachin A.L. Pattern analysis: ordinal-invariant pattern-clustering. *Control sciences*. 2016, N 4, P. 2–9. (In Russ.)
- Savoia A., Sen K. *Measurement and evolution of state capacity: Exploring a lesser known aspect of governance*. Manchester, UK, 2012, 28 p. (ESID Working Paper 10)
- Shawe-Taylor J., Cristianini N. *Kernel methods for pattern analysis*. Cambridge: Cambridge univ. press, 2004. Mode of access: http://read.pudn.com/downloads190/ebook/893343/Kernel_Methods_for_Pattern_Analysis/0521813972.pdf (accessed: 20.04.2019.)
- Wang S. The Rise of the Regions: Fiscal Reform and the Decline of Central State Capacity in China. In: *The Waning of the Communist State: Economic Origins of Political Decline in China and Hungary*. Ed. by A. Walder. Berkeley, Univ. of California press, 1995, P. 87–114.
- Xu R., Wunsch D.C. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. 2005, Vol. 16, N 3, P. 645–678. <https://doi.org/10.1109/TNN.2005.845141>