Санкт-Петербургский государственный университет
Кафедра математической теории игр и статистических решений

Староверова Ксения Юрьевна

Магистерская диссертация

Кластеризация временных рядов

Направление 01.04.02

Прикладная математика и информатика

Магистерская программа «Исследование операций и системный анализ»

Научный руководитель,
доктор технических наук,
профессор
Буре В. М.

Санкт-Петербург
2017

Оглавление

Введение

В связи с ростом вычислительных мощностей и возможностей собирать и хранить данные, одна из наиболее актуальных областей современного мира — это анализ данных. Активное развитие получили методы машинного обучения, так как они способны «учиться» на тестовых данных и призваны решать самые разнообразные задачи: классификация данных, прогнозирование (в частности, построение регрессионной модели), кластеризация, поиск аномалий и др. Люди всегда стремились предсказывать будущее или поведение других людей по имеющимся данным, особенно актуальна эта задача в области страхования и финансов. Однако сегодня методы машинного обучения используются в диагностике заболеваний, при изучении био-физических особенностей организма, в распознавании изображений и речи, автоматизации труда и т.д.

Наиболее сложными являются методы, которые работают «без учителя». В таких задачах нет тестовой выборки, по которой можно подобрать параметры или проверить результат. Кластеризация данных относится именно к этой группе. Задача состоит в распределении объектов (данных) по группам таким образом, чтобы внутри каждой группы оказались объекты, обладающие высокой степенью сходства в некотором отношении, которое является принципиально важным для рассматриваемой задачи, а между различными группами обнаруживались бы существенные различия. Для выделения кластеров используются только сами данные, количество кластеров является неизвестной величиной. Процедура распределения по кластерам (подход) может проходить по разному. Здесь выделяют вероятностные, теоретико-графовые, иерархические, нечеткие подходы, эта классификация условная, существуют и другие алгоритмы. Результат будет зависеть не только от выбора подхода, но и от способа определения количества кластеров и выбора метрики.

Работа посвящена кластерному анализу временных рядов, который выделен из общей задачи кластеризации в связи с тем, что данные зависят от времени. Это требует подбора специальных метрик, учитывающих временные особенности. Выбор темы обусловлен не только ее актуальностью, но и наличием практической задачи, рассмотрение которой также приводится в работе. Постановка проблемы и данные предоставлены медицинским информационно-аналитическим центром.

В области кластеризации временных рядов была проделана большая работа: описано применение алгоритма динамической трансформации шкалы в кластерном анализе [1], [2]; этот же алгоритм, но с помощью теории скрытых марковских моделей предложен в [3]; использование коэффициентов автокорреляции, спектральных характеристик, вейвлет-коэффициентов отмечено в работах [4], [5], более полный обзор существующих методов представлен в статье [6].

В первой главе дан обзор существующих метрик и приведено описание новой метрики, разработанной для кластеризации коротких временных рядов. Во второй главе показан кластерный анализ районов Санкт-Петербурга по показателю заболеваемости с 1999 по 2014 гг., получено несколько моделей, по которым построены стабильные кластеры. В том числе, рассмотрен многомерный анализ. В третьей главе рассмотрен метод работы с объектами, которые потенциально могут относиться к нескольким кластерам, работа алгоритма показана на примере детской заболеваемости.

Глава 1. Расстояние, основанное на характеристиках временных рядов

В современном мире для решения многих задач используются методы машинного обучения [7], к таким задачам относятся кластеризация и классификация данных. Большую роль при проведении кластеризации и классификации объектов играет выбор меры сходства или различия объектов для построения матрицы расстояний, так как неточно подобранная мера повлечет за собой неточные результаты решения задачи. Несмотря на то что область анализа данных и машинного обучения развивается довольно стремительно, мы столкнулись с проблемой выбора меры различия временных рядов, которую можно было бы использовать для кластеризации районов Санкт-Петербурга. Задача кластеризации временных рядов выделена отдельно в связи с тем, что меры близости должны учитывать характер изменчивости данных. Таким образом, если использовать классические метрики, например, Евклидову, то расстояние не изменится при перестановке наблюдений местами, что является некорректным и противоречит природе временных рядов.

Существующие меры различия дают хорошие результаты при работе с длинными рядами и зачастую даже применяют различные приемы по сокращению размерности, но при кластеризации коротких временных рядов такие меры могут давать не достаточно точные оценки близости объектов, что влечет увеличение количества ошибок кластеризации. В этой главе предложена мера различия временных рядов «мера различия, основанная на характеристиках» (MPOX) или «characteristics based distance» (CBD) [8], которая учитывает различие между статистическими характеристиками объектов кластеризации или классификации. Эксперименты на искусственных наборах данных показали, что выбор MPOX для кластери-

зации временных рядов оправдан, так как количество ошибок кластеризации оказалось меньше, чем при использовании многих других наиболее известных мер различия, время работы алгоритма оказалось также меньше. Построение матрицы MPOX-расстояний реализовано на языке R, в этом же пакете проведен сравнительный анализ данного метода с другими для этого с использованием библиотек «cluster» и «TSclust».

1.1. Обзор методов библиотеки «TSclust» пакета R

Библиотека представляет широкий выбор мер различия для кластерного анализа временных рядов и несколько наборов искусственных и реальных данных [9], [10], поэтому она выбрана для проведения экспериментов и сравнительного анализа существующих мер с MPOX. Построение матрицы расстояний производится при помощи функции $diss(SERIES, METHOD, \dots)$, где $SERIES$ — это матрица, список или датафрейм с набором временных рядов, а $METHOD$ — строка с названием меры различия.

Существует условное разделение методов на 3 группы. Методы *свободные от предположений о модели* строят расстояние непосредственно между наблюдениями двух рядов или некоторыми их характеристиками.

$EUCL$ евклидово расстояние, широко применяется при кластеризации, но не всегда может быть использовано, если объектом является временной ряд, в силу того что зависимость наблюдений от времени игнорируется.

ACF расстояние между коэффициентами автокорреляции рядов.

$PACF$ расстояние между частичными коэффициентами автокоррелиции рядов.

COR мера основана на вычислении корреляции Пирсона между наблюдениями объектов.

CORT в литературе носит название «адаптивный индекс различия» (adaptive dissimilarity index), адаптивность связана с возможностью регулировать влияние на результат двух составляющих: различия с точки зрения значений наблюдений и различия с точки зрения динамики временных рядов.

FRECHET сужение расстояния Фреше на дискретный случай. Расстояние Фреше было предложено для вычисления расстояния между кривыми, поэтому оно подходит для вычисления расстояния между временными рядами, однако оно не принимает во внимание поведение временных рядов или характер их изменчивости.

DTWARP алгоритм динамической трансформации шкалы, широко применяется при классификации и кластеризации речи, так как инвариантен относительно масштабирования и сдвига по временной оси, имеет тот же недостаток, что и предыдущий метод, кроме того, использование данного метода на любых данных неоправдано, так как иногда реальные различия в поведении могут быть неучтены или значительно уменьшены после преобразований временной оси.

PER расстояние между периодограммами.

INT.PER интегрирование периодограм перед вычислением расстояния имеет преимущества. Это расстояние может быть нормировано, в таком случае большее значение имеет форма кривой временного ряда.

SPEC.GLK различие между двумя рядами оценивается в терминах значения статистики равенства логарифма спектров, для проверки гипотезы изпользуется обобщенный критерий отношения правдоподобия [11].

SPEC.ISD вычисляется как интеграл от квадратов разностей между непараметрическими оценками логарифма спектра.

SPEC.LLR представлено отношением локальных линейных спектральных оценок.

Другая группа —*методы основанные на предположениях о модели.* Если предположить, что производится кластеризация стационарных временных рядов, тогда их можно описать достаточно точно моделью ARMA, поэтому можно говорить о различии временных рядов с точки зрения подобранных моделей [4], [5]. Модели подбираются согласно критерию AIC.

AR.MAH строится статистика, отражающая значимость различий процесов, для этого используются коэффициенты авторегрессии, дисперсия белого шума, матрицы ковариаций рядов. Статистика распределена как χ^2, нулевая гипотеза говорит о абсолютном сходстве процессов.

AR.LPC.CEPS расстояние между ЛПК-кепстральными коэффициентами, т. е. кепстральными коэффициентами сигнала, которые вычисляются через коэффициенты авторегриссии.

AR.PIC так как считается, что любой стационарный процесс можно аппроксимировать моделью AR(∞), этот метод находит евклидово расстояние между коэффициентами модели подобранной по критериям AIC или BIC.

Методы основанные на оценке сложности временного ряда отличаются от предыдущих тем, что они не учитывают временные характеристики и особенности процессов, не выдвигают предположения о модели, которой можно аппроксимировать процесс, зато оценивают уровень информации, которую содержит каждый временной ряд.

Важно отметить, что каждый метод обладает своей спецификой. Во-первых, выбор одной из трех групп должен быть не случайным и соответствовать целям исследования, во-вторых выбор меры также должен

CDM — мера, основанная на различии физического размера сжатых временных рядов. Допускается выбор одного из трех алгоритмов сжатия: «gzip», «bzip2», «xz». Возможно использование символьного представления процесса, например SAX-методов.

NCD — отличается от предыдущего метода только формулой вычисления расстояния (происходит нормализация) [12].

CID — модификация евклидова расстояния: происходит домножение на коэффициент, определяющий отношение сложностей двух рядов.

MINDIST.SAX — расстояние находится как минимум после некоторых преобразований: нормализация временного ряда, сокращение размерности алгоритмом PAA, преобразование алгоритмом SAX в символьный ряд.

PDC — сначала выполняется перестановка наблюдений в восходящем порядке, после чего строится код, отражающий распределение упорядоченных наблюдений на первоначальном ряде, этот код используется для оценки сложности процесса. Расстояние между двумя рядами вычисляется как различие между построенными кодами.

быть обусловлен спецификой задачи. Например, если необходимо производить кластеризацию длинных рядов ежесекундно, при этом визуализировать данные, а точностью можно пренебречь, то стоит отдать предпочтение более простым методам, таким как вычисление евклидова расстояния или коэффициентов корреляции Пирсона, и наоборот, если проводится исследование, в котором важна точность результатов, то выбирать нужно более сложные методы, например, динамическое преобразование шкалы, методы, основанные на спектре процесса, преобразованиях вейвлетами.

1.2. Описание метода

Создание этого метода началось с формулировки недостатка существующих методов: многие из них не предназначены для работы с короткими рядами [8]. Вопрос кластеризации временных рядов малой размерности является актуальным, так как статистика по многим экономическим, социальным, демографическим и другим показателям собирается нечасто — один раз в год, квартал или месяц и доступна за небольшие периоды в несколько лет. Именно поэтому перед нами стоял вопрос: каким образом можно представить информацию о временном ряде, кроме уже существующих способов, так, чтобы эта информация являлась достаточно точным описанием и для коротких временных рядов.

Метод основан на вычислении характеристик временного ряда, которые делятся на три группы: не зависящие от времени, описывающие динамику и изменчивость. Для каждой группы характеристик вычисляется величина, отражающая различие временных рядов по характеристикам из конкретной группы, а затем строится линейная комбинация этих величин. Представим имеющиеся временные ряды в виде матрицы, где строки соответствуют временным рядам. Таким образом матрица $M = [n \times T]$ задает n временных рядов с T наблюдениями. Введем обозначение временного ряда $M_i = \{m_{i,1}, m_{i,2}, \ldots, m_{i,T}\}$, соответствующего i-ой строке матрицы. Тогда расстояние между двумя рядами вычисляется по формуле

$$\begin{aligned} dist_{CBD}(M_{i_1}, M_{i_2}) = \alpha\, dist_1(M_{i_1}, M_{i_2}) + \beta\, dist_2(M_{i_1}, M_{i_2}) + \\ + (1 - \alpha - \beta)\, dist_3(M_{i_1}, M_{i_2}), \end{aligned} \tag{1.1}$$

коэффициент $\alpha \in [0, 1]$ отвечает за влияние расстояния между характеристиками, не зависящими от времени, на итоговый результат; коэффициент $\beta \in [0, 1]$ — за влияние расстояния между характеристиками, отражающими динамику временного ряда; аналогично $(1 - \alpha - \beta)$ отвечает за влияние расстояния между характеристиками изменчивости ряда. Кажде слагаемое принимает значения на отрезке $[0, 1]$, что достигается нормированием данных.

Перед вычислением характеристик, *не зависящих от времени*, пре-

образуем временные ряды:

$$\overline{m}_{i,t} = \frac{m_{i,t} - \min M}{\max M - \min M}, \tag{1.2}$$

$$\overline{M} = \{m_{i,t}\}, i = \overline{1,n}, t = \overline{1,T}. \tag{1.3}$$

После этого по матрице $\overline{M}$ для каждого временного ряда (строки) вычисляем матрицу характеристик $C = [n \times 5]$: $\{c_{1,1}, c_{2,1}, \ldots, c_{n,1}\}^T$ — столбец средних значений каждого ряда, $\{c_{1,2}, c_{2,2}, \ldots, c_{n,2}\}^T$ — стандартных отклонений, $\{c_{1,3}, c_{2,3}, \ldots, c_{n,3}\}^T$ — медиан, $\{c_{1,4}, c_{2,4}, \ldots, c_{n,4}\}^T$ — минимумов, $\{c_{1,5}, c_{2,5}, \ldots, c_{n,5}\}^T$ — максимумов. Здесь не учитывается зависимость данных от времени. Расстояние между временными рядами с точки зрения значений наблюдений вычисляется по формуле:

$$dist_1(M_{i_1}, M_{i_2}) = \sqrt{\frac{1}{5}\sum_{k=1}^{5}(c_{i_1,k} - c_{i_2,k})^2} \tag{1.4}$$

Для вычисления *различий в динамике* для каждого временного ряда строим четыре вектора, характеризующих его поведение. Определим первые разности ряда с лагом l

$$FD_{i,t}(l) = M_{i,t+l} - M_{i,t}, \tag{1.5}$$

где $i = \overline{1,n},\ t = \overline{1,T-l}$. Для описания динамики временного ряда будем использовать первые разности с лагом 1

$$FD_{i,t} = FD_{i,t}(1). \tag{1.6}$$

Определим построение первой матрицы $D^{(1)} = \{d^{(1)}_{i,t}\}$, $i = \overline{1,n}$, $t = \overline{1,T-1}$. Каждая строка $D^{(1)}$ отражает динамику процесса, а именно, показывает участки неубывания

$$d^{(1)}_{i,t} = \begin{cases} 1, \text{ если } FD_{i,t} \geq 0, \\ 0, \text{ иначе.} \end{cases} \tag{1.7}$$

Вторая матрица $D^{(2)} = \{d_{i,t}^{(2)}\}$, $i = \overline{1,n},\ t = \overline{1,T}$ показывает как временной ряд флуктуирует около среднего значения

$$d_{i,t}^{(2)} = \begin{cases} 1, \text{если } M_{i,t} \geq \mathrm{E}(M_i), \\ 0, \text{иначе.} \end{cases} \tag{1.8}$$

Третья и четвертая матрицы $D^{(3)} = \{d_{i,t}^{(3)}\}$, $D^{(4)} = \{d_{i,t}^{(4)}\}$ $i = \overline{1,n},\ t = \overline{1,T}$ показывают наблюдения, которые отклонились от среднего значения больше чем на стандартное отклонение

$$d_{i,t}^{(3)} = \begin{cases} 1, \text{если } M_{i,t} \geq \mathrm{E}(M_i) + \sqrt{\mathrm{D}(M_i)}, \\ 0, \text{иначе,} \end{cases} \tag{1.9}$$

$$d_{i,t}^{(4)} = \begin{cases} 1, \text{если } M_{i,t} \leq \mathrm{E}(M_i) - \sqrt{\mathrm{D}(M_i)}, \\ 0, \text{иначе,} \end{cases} \tag{1.10}$$

где $\mathrm{E}(M_i)$ — среднее значение временного ряда i, $\mathrm{D}(M_i)$ — его дисперсия. Различия в динамике временных рядов M_{i_1} и M_{i_2} вычисляются по следующей формуле

$$dist_2(M_{i_1}, M_{i_2}) = \frac{1}{4}\left(\frac{1}{T-1}\sum_{t=1}^{T-1} d_{i_1,t}^{(1)} \oplus d_{i_2,t}^{(1)} + \sum_{k=2}^{4}\frac{1}{T}\sum_{t=1}^{T} d_{i_1,t}^{(k)} \oplus d_{i_2,t}^{(k)}\right), \tag{1.11}$$

где $d_{i_1,t}^{(k)} \oplus d_{i_2,t}^{(k)}$ — покомпонентное сложение по модулю 2.

Характеристики, отвечающие за *изменчивость* включают в себя 15 величин, которые вычисляются для первых разностей ряда с лагом 1 и 2, их можно представить в виде матрицы $V = [n \times 15]$. Далее будем обозначать столбец с номером k матрицы V как $V_{*,k} = \{v_{1,k}, v_{2,k}, \ldots, v_{n,k}\}$, где $k = \overline{1,15}$. Для вычисления компонент матрицы V сначала выполним преобразование матрицы M, таким образом, чтобы максимальному значению каждого ряда соответствовала 1, а минимальному — 0

$$\widetilde{m}_{i,t} = \frac{m_{i,t} - \min M_i}{\max M_i - \min M_i}, \tag{1.12}$$

$$\widetilde{M} = \{\widetilde{m}_{i,t}\}, i = \overline{1,n},\ t = \overline{1,T}. \tag{1.13}$$

Преобразование (1.12)–(1.13) позволяет вычислять характеристики временных рядов, принимая во внимание только их изменчивость, а не значения наблюдений. Тогда первые разности для преобразованных рядов задаются формулой

$$\widetilde{FD}_{i,t}(l) = \widetilde{M}_{i,t+l} - \widetilde{M}_{i,t},\ i = \overline{1,n},\ t = \overline{1,T-l}. \tag{1.14}$$

Определим компоненты первых трех столбцов матрицы V: среднюю скорость роста $V_{*,1}$, спада $V_{*,2}$ и изменения ряда $V_{*,3}$. Вычислим количество участков возрастания и убывания ряда при $l = 1$

$$N_i^+(l) = \sum_{t=1}^{m-l} \mathrm{I}\{\widetilde{FD}_{k,t}(l) > 0\}, \tag{1.15}$$

$$N_i^-(l) = \sum_{t=1}^{m-l} \mathrm{I}\{\widetilde{FD}_{k,t}(l) < 0\}, \tag{1.16}$$

где I — это индикатор (принимает значение 1, когда верно выражение в скобках, иначе — 0). Тогда компоненты столбцов $V_{*,1},\ V_{*,2},\ V_{*,3}$ задаются как

$$V_{i,1}(l) = \frac{1}{N_i^+(l)} \sum_{t=1}^{T-l} \widetilde{FD}_{i,t}(l)\ \mathrm{I}\{\widetilde{FD}_{i,t}(l) > 0\}, \tag{1.17}$$

$$V_{i,2}(l) = \frac{1}{N_i^-(l)} \sum_{t=1}^{T-l} \widetilde{FD}_{i,t}(l)\ \mathrm{I}\{\widetilde{FD}_{i,t}(l) < 0\}, \tag{1.18}$$

$$V_{i,3} = \mathrm{E}(\widetilde{FD}_{i,*}(l)), \tag{1.19}$$

где $i = \overline{1,n},\ l = 1$. Таким образом по формулам (1.15)-(1.19) при $l = 1$ вычисляются $V_{*,1},\ V_{*,2},\ V_{*,3}$, если положить $l = 2$ и вновь воспользоваться формулами (1.15)-(1.19), получим компоненты столбцов $V_{*,4},\ V_{*,5},\ V_{*,6}$.

Следующие характеристики показывают величину наибольшего роста и наибольшего спада, процесс их вычисления приведен в виде алгоритма ниже.

Вход: FD — матрица первых разностей, где строка с номером i соответствует первой разности временного ряда i;

Выход: для каждого ряда с номером i пара значений $\{Grouth[i], Decline[i]\}$;

$n \leftarrow$ количество строк FD

$T \leftarrow$ количество столбцов FD

Для i **от** 1 **до** n

$Grouth[\, i \,] \leftarrow 0$

$Decline[\, i \,] \leftarrow 0$

$Sum \leftarrow FD[\, i,\, 1 \,]$

Для t **от** 2 **до** T

Если $(FD[\, i,\, t \,] \cdot FD[\, i-1,\, t \,]) < 0$ **И** $Sum > Grouth[\, i \,]$

То $Grouth[\, i \,] \leftarrow Sum$

$Sum \leftarrow FD[\, i,\, t \,]$

Если $(FD[\, i,\, t \,] \cdot FD[\, i-1,\, t \,]) < 0$ **И** $Sum < Decline[\, i \,]$

То $Decline[\, i \,] \leftarrow Sum$

$Sum \leftarrow FD[\, i,\, t \,]$

Если $(FD[\, i,\, t \,] \cdot FD[\, i-1,\, t \,]) > 0$

То $Sum \leftarrow Sum + FD[\, i,\, t \,]$

Столбцы $V_{*,7},\ V_{*,8}$ вычисляются по описанному выше алгоритму для первых разностей ряда с лагом 1, а столбцы $V_{*,9},\ V_{*,10}$ — для первых разностей ряда с лагом 2. Последние 5 столбцов матрицы V соответствуют минимальному значению, квартилям и максимальному значению первых разностей с лагом 1.

Различия временных рядов M_{i_1} и M_{i_2} по отношению к их изменчивости вычисляется согласно формуле

$$dist_3(M_{i_1}, M_{i_2}) = \sqrt{\frac{1}{15}\sum_{k=1}^{15}(v_{i_1,k} - v_{i_2,k})^2}. \tag{1.20}$$

Таким образом расстояние, основанное на характеристиках временного ряда полностью определено формулами (1.1), (1.4), (1.11), (1.20). Влияние различия временных рядов по отношению к значению наблюдений, динамике и изменчивости регулируется с помощью коэффициентов $\alpha \in [0, 1]$ и $\beta \in [0, 1]$. Расстояние, введенное в формуле (1.1) является метрикой.

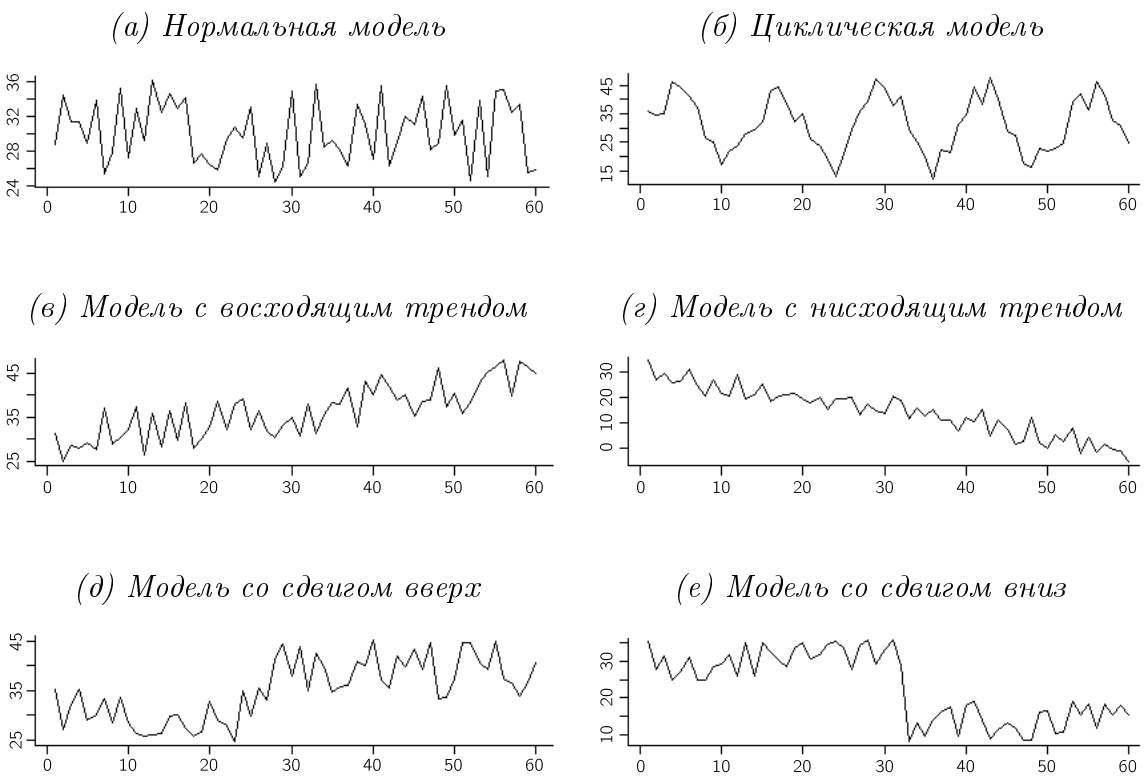


Рис. 1.1. Пример реализаций разных моделей, по оси абсцисс — время, по оси ординат — значение временного ряда.

1.3. Эксперимент

Идея использовать характеристики временных рядов для проведения кластеризации уже была описана в [13], однако удовлетворительных результатов получено не было. Проведен эксперимент на том же наборе искусственных данных, который содержит 600 временных рядов, сгенерированных по 6 следующим моделям $y(t), t = \overline{1, 60}$:

1) нормальная модель: $y(t) = m + rs$, где $m = 30, s = 2, r \in [-3, 3]$;

2) циклическая модель: $y(t) = m + rs + a \sin \frac{2\pi t}{T}$, где $a, T \in [10, 15]$;

3) модель с восходящим трендом: $y(t) = m + rs + gt$, где $g \in [0.2, 0.5]$;

4) модель с нисходящим трендом: $y(t) = m + rs - gt$, где $g \in [0.2, 0.5]$;

5) модель со сдвигом вверх: $y(t) = m + rs + kx$, где $x \in [7.5, 20], k = 0$ до момента $\hat{t}$ и $k = 1$ после этого момента, $\hat{t} \in [20, 40]$;

6) модель со сдвигом вниз: $y(t) = m + rs - kx$, где $x \in [7.5, 20], k = 0$ до момента $\hat{t}$ и $k = 1$ после этого момента, $\hat{t} \in [20, 40]$.

Теперь стоит задача классификации, а не кластеризации, так как количество кластеров в искусственных данных нам известно. Однако, в реальных задачах меры различия временных рядов могут быть использованы и в кластеризации, и в классификации.

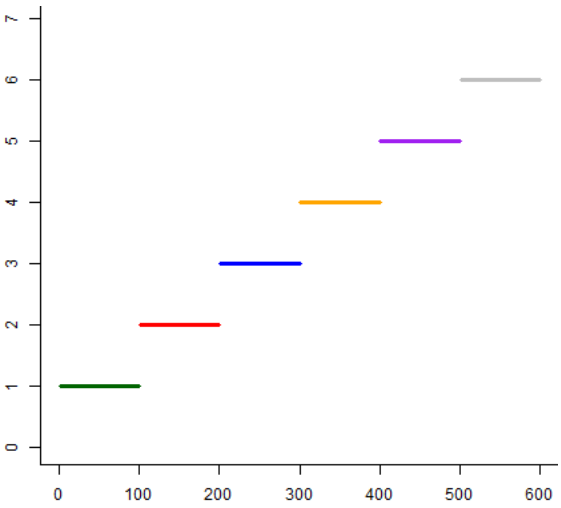
Примеры временных рядов каждой модели представлены на рис. 1.1(а-. Сравнивать метод *CBD* с другими расстояниями можно по проценту правильно классифицированных объектов P. Например, в [13] этот показатель составил 47.3%. При классификации методом *CBD* с коэффициентами $\alpha = 0.55, \beta = 0.1$ эффективность кластеризации P составила 85.2%. Такие коэффициенты выбраны, как доставляющие максимум среднего значения индекса силуэта кластеров. Однако при $\alpha = 0.4, \beta = 0.4$ величина P составляет уже более 91 %. На рис.1.2(а) показано как должны быть распределены объекты по кластерам, а на рис.1.2(б) — результаты, полученные нашим методом с коэффициентами $\alpha = 0.55, \beta = 0.1$.

Сравнение CBD с методами библиотеки «TSclust» пакета R показало, что для данного набора данных наш метод дает лучшие результаты. Для большей части методов $P < 60\%$ и только для алгоритма динамической трансформации шкалы (DTW - dynamic time warping) $P = 84.7\%$, однако время построения матрицы расстояний методом CBM составляет 16.04 с., а для DTW на том же компьютере время построения составило 445.68 с. Распределение объектов по кластерам для этого метода показано на рис.1.2(в).

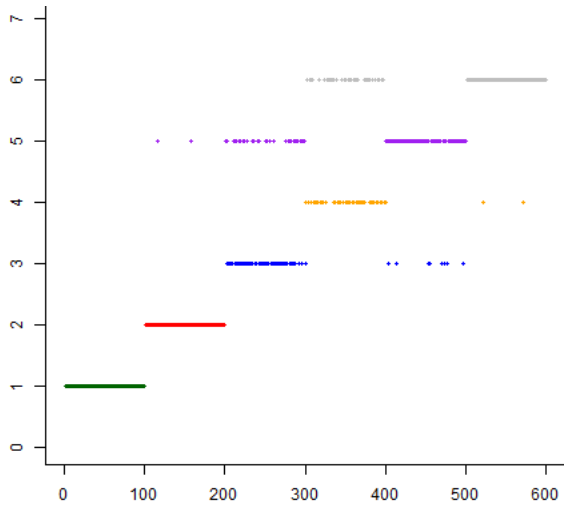
Также можно сравнить индексы оценки силуэта кластеров, данная величина часто применяется для оценки качества кластеризации и в нашем случае может говорить о некоторой устойчивости, так как высокий индекс оценки силуэта говорит о том, что расстояние между объектами внутри кластера мало, а между элементами соседних кластеров — велико [14]. Индекс оценки силуэта может принимать значения от 0 до 1, чем ближе значение к 1, тем лучше произведена кластеризация. Можно заметить по табл. 1.1, что в среднем индекс выше при кластеризации методом CBM.

Второй эксперимент проведен с искусственными данными библиотеки «TSclus». Набор содержит по 3 реализации длиной 200 каждой из 6 следующих моделей:

(а) Распределение объектов по кластерам согласно моделям



(б) Кластеризация методом CBM



(в) Кластеризация методом DTW

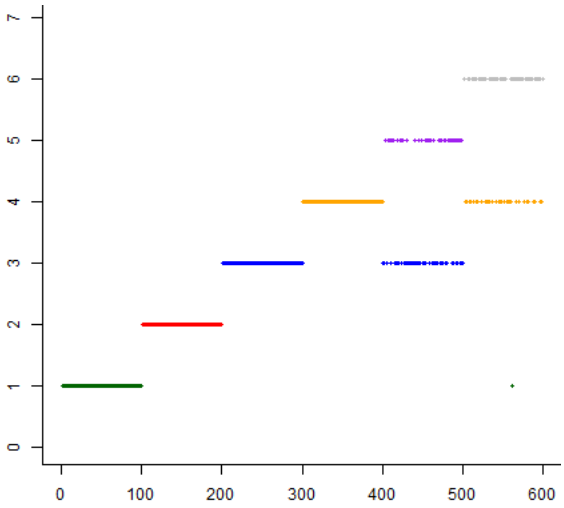


Рис. 1.2. Распределение объектов по кластерам, по оси абсцисс — номер объекта, по оси ординат — номер кластера.

1) $X_t = 0.6X_{t-1} + \epsilon_t$;

2) $X_t = (0.3 - 0.2\epsilon_{t-1})X_{t-1} + 1 + \epsilon_t$;

3) $X_t = (0.9\exp(-X_{t-1}^2) - 0.6)X_{t-1} + 1 + \epsilon_t$;

4) $X_t = (0.3X_{t-1} + 1)I(X_{t-1} \geq 0.2) - (0.3X_{t-1} - 1)I(X_{t-1} < 0.2) + \epsilon_t$;

5) $X_t = 0.7|X_{t-1}|(2 + |X_{t-1}|)^{-1} + \epsilon_t$;

6) $X_t = 0.8X_{t-1} - 0.8X_{t-1}(1 + \exp(-10X_{t-1}))^{-1} + \epsilon_t$;

Эксперименты на этих данных уже проводились в [10], наилучшие результаты показали методы «INT.PER» $P = 88.9\%$ и «SPEC.LLR» $P = 83.3\%$.

Таблица 1.1. **Индекс оценки силуэтов для классификаций методами CBD и DTW**

Метод	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5	Кластер 6
CBD	0.39	0.56	0.13	0.42	0.50	0.13
DTW	0.58	0.22	0.14	0.12	0.28	0.25

Наш метод показывает результаты немного хуже, чем предыдущие 2 метода, однако лучше, чем остальные: $P = 82.8\%$ при $\alpha = 0.3, \beta = 0.3$.

В [15] был представлен набор данных для валидации методов кластеризации. База содержит 18 пар временных рядов (*Приложение 1*), которые представляют разные предметные области. Например, ЭКГ грудной и брюшной полости плода, спрос на электроэнергию за различные периоды одного года и т.д. Авторы [15] предлагают проводить анализ следующим образом: по матрице расстояний построить дендрограмму. Тогда согласно теоретическому распределению, пары рядов из одной области будут находиться в одном поддереве на последнем уровне. Для оценки качества предлагается рассчитать величину Q, которая показывает относительное количество правильно определенных пар. Согласно проведенным экспериментам в 2004 году более $3/4$ существующих методов показали наихудший результат, т. е. $Q = 0$. Наилучшие результаты здесь показывают методы, основанные на символьных преобразованиях. Новый метод корректно определил 6 пар, $Q = 0.33$ рис. 1.3.

Такой результат не является удовлетворительным для реальной задачи, но если расценивать параметр Q исключительно как показатель для сравнения различных методов кластеризации временных рядов, то можно сказать, что MPOX превосходит бóльшую часть существующих расстояний. Более того, можно заметить, что временные ряды $1, 2, 3, 4$ определены в одну группу и их, действительно, можно считать сильно похожими (*Приложение 1*), как и ряды $21, 22, 23, 24$ и $29, 30, 35, 36$.

Авторы [16] собали большую базу временных рядов для задачи классификации, для тестирования метода MPOX выбран один из наборов, в котором представлены данные из Центра землетрясений Южной Калифорнии с 1 декабря 1967 по 2003 гг. Представлено 139 временных рядов длиной 512 наблюдений. Эксперимент показал наилучший результат, полученный

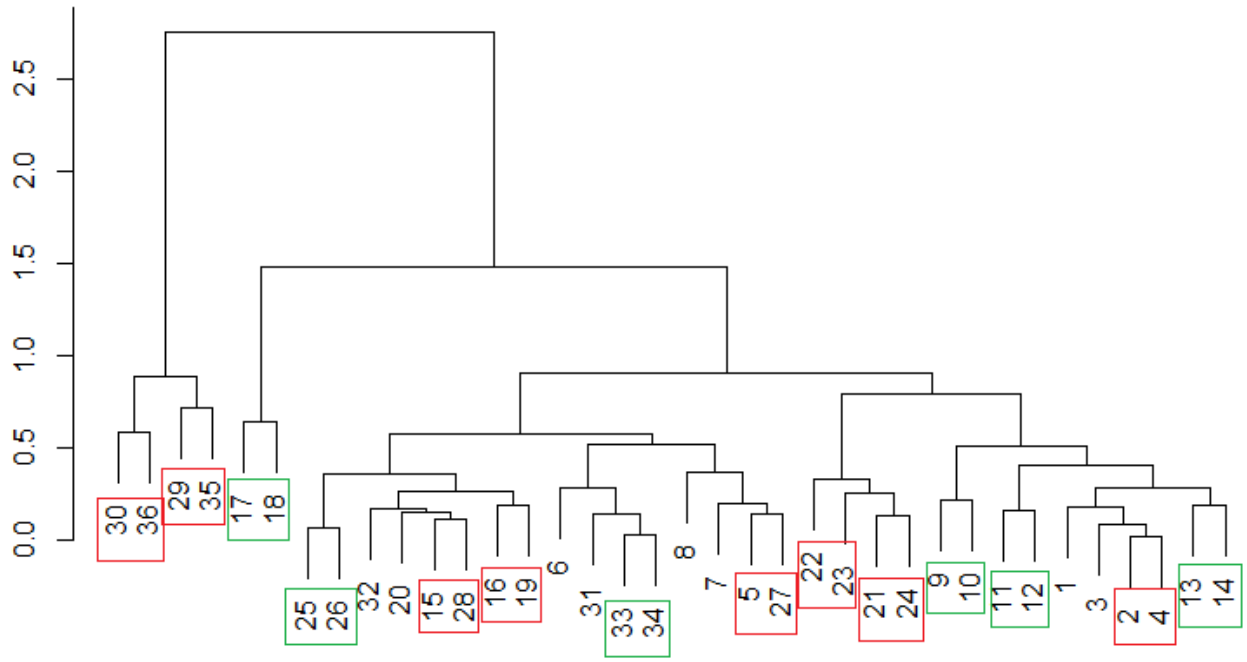


Рис. 1.3. Дендрограмма временных рядов, описывающих различные предметные области

методом MPOX — 53 %, метод, основанный на спектральной плотности «SPEC.LLR» — 52%, остальные методы правильно распознали менее 50 % объектов.

1.4. Выводы

Метод, основанный на вычислении характеристик, показал неплохие результаты при проведении экспериментов, что позволяет его использовать в задачах кластерного анализа временных рядов. Преимуществами являются простота и скорость работы алгоритма. Стоит также отметить, что данный метод учитывает и динамику временной ряда и характер изменчивости, при этом предложенная мера различия является метрикой. Недостаток алгоритма заключается в подборе параметров α и β, поэтому естественным развитием становится разработка критериев по выбору этих коэффициентов. Однако постановка некоторых задач позволяет понять, какие характеристики оказываются более важными для кластеризации, в таком случае трудности с подбором коэффицентов не возникают.

Глава 2. Задача кластеризации районов Санкт-Петербурга по показателю заболеваемости

Система здравоохранения в России устроена таким образом, что контролирующие органы имеют иерархическую структуру. Например, в Санкт-Петербурге существует Комитет по здравоохранению, который является ответственным за город, ему подчиняются районные комитеты, которые в свою очередь контролируют учреждения, относящиеся только к ним. Таким образом статистика проделывает долгий путь от обычного врача-статиста до сотрудников Комитета по здравоохранению или Министерства по здравоохранению. В связи с этим возникает вопрос: есть ли статистически значимые различия между районами города с точки зрения показателей здравоохранения. Если они есть, то они могут свидетельствовать о сборе некачественной статитики в районе, о плохой работе учреждений и управляющих органов и о других факторах, влияющих на население района.

Имеются данные с 1999 по 2014 гг. по показателю заболеваемости для каждого из 18 районов Санкт-Петербурга. Показатель заболеваемости рассчитан как отношение случаев (форма 12 федерального статистического учета) к численности населения. Причем он вычислен для трех возрастных групп (различных с точки зрения здравоохранения): дети в возрасте 0–14 лет, подростки 15–17 лет и взрослые (старше 18 лет). Необходимо выяснить есть ли статистически значимые различия между районами по этому показателю [17].

Естественным способом выявления неоднородности объектов является применение методов кластерного анализа [18], так как задача кластеризации состоит в разбиении базы данных на группы таким образом, что

внутри группы объекты более схожи, чем объекты из разных групп. Однако, для проведения кластеризации районов, нужно построить матрицу расстояний между ними, в этом состоит основная сложность. Она связана с необходимостью учитывать тот факт, что объектами кластеризации являются районы, которые представлены многомерными временными рядами [19].

2.1. Основные понятия

В этом параграфе кратко рассмотрены основные понятия, которые понадобятся для проведения исследования.

Определение. *Многомерный временной ряд* — это набор последовательных наблюдений $X(t) \in R^n$, где $t \in \{1, 2, \ldots, T\}$ обозначает момент времени, в который выполнено наблюдение. Таким образом в момент времени $t = \overline{1,T}$ выполняется измерение n характеристик. Соответственно, в *одномерном временном ряде* $n = 1$.

Определение. *Длина временного ряда* — это количество наблюдений T.

Наиболее часто для нахождения расстояния между объектами используют *евклидову метрику*, если $X(t)$ и $Y(t)$ — одномерные временные ряды длиной T, то

$$d_E(X,Y) = \sqrt{\big(X(1) - Y(1)\big)^2 + \cdots + \big(X(T) - Y(T)\big)^2} \tag{2.1}$$

определяет евклидово расстояние между ними.

Расстояние Фреше между временными рядами $X(t), Y(t), t = \overline{1,T}$ определяется как

$$d_F(X,Y) = \min_{r \in M} \max_{j=1,\ldots,k} |X(a_j) - Y(b_j)|, \tag{2.2}$$

где M — это множество, которое содержит всевозможные последовательности пар, не нарушающих порядок наблюдений,

$$r = \Big(\big(X(a_1), Y(b_1)\big), \ldots, \big(X(a_k), Y(b_k)\big)\Big), r \in M,$$

пары $\big(X(a_j), Y(b_j)\big) \in r$ удовлетворяют следующим 5 свойствам:

1) $a_1, \ldots, a_k, b_1, \ldots, b_k \in \{1, \ldots, T\}$;

2) $a_1 = b_1 = 1$;

3) $a_k = b_k = T$;

4) $a_{j+1} = a_j$ или $a_{j+1} = a_j + 1$;

5) $b_{j+1} = b_j$ или $b_{j+1} = b_j + 1$.

Формула (2.2) является сужением на дискретный случай меры, предложенной французским математиком Maurice René Fréchet в 1906 году для нахождения расстояния между кривыми. Интуитивное описание дали T. Eiter и H. Mannila в [20].

Регулируемый индекс различия (an adaptive dissimilarity index) [21] — это мера различия временных рядов, которая учитывает и характер изменчивости, и значения наблюдений

$$\mathrm{CorT}(X, Y) = \frac{\sum\limits_{t=1}^{T-1} \big(X(t+1) - X(t)\big)\big(Y(t+1) - Y(t)\big)}{\sqrt{\sum\limits_{t=1}^{T-1} \big(X(t+1) - X(t)\big)^2} \sqrt{\sum\limits_{t=1}^{T-1} \big(Y(t+1) - Y(t)\big)^2}}, \quad (2.3)$$

$$d_{\mathrm{CorT}}(X, Y) = f\big(\mathrm{CorT}(X, Y)\big)\delta_{conv}(X, Y), \quad (2.4)$$

$$f(\alpha) = \frac{2}{1 + \exp k\alpha}, \quad (2.5)$$

$k > 0$ — параметр. Значение индекса (2.4) зависит от двух величин: различие в поведении рядов представлено формулой (2.3), а расстояние между наблюдениями $\delta_{conv}(X, Y)$ вычисляется любым традиционным методом (метрика Евклида, Минковского, Манхэттенская и т.д.). Параметр k регулирует влияние компонент на итоговый результат.

Расстояние Эрос (Eros) [22] — мера различия многомерных временных рядов, основанная на сингулярном разложении матриц ковариаций рядов. Представим многомерные временные ряды $X(t)$ и $Y(t)$ в виде матриц X и Y порядка $n \times T$. Для них построим матрицы ковариаций переменных временных рядов M_x и M_y порядка $n \times n$. Выполним сингулярное

разложение матриц ковариаций

$$M_x = U_x \Sigma_x V_x^T, \tag{2.6}$$

$$M_y = U_y \Sigma_y V_y^T. \tag{2.7}$$

Для вычисления схожести рядов $X(t)$ и $Y(t)$ используются матрицы правых сингулярных векторов $V_x = \{v_{x_1}, v_{x_2}, \ldots, v_{x_n}\}$

$$\mathrm{Eros}(V_x, V_y, w) = \sum_{i=1}^{n} w_i |\langle v_{x_i}, v_{y_i} \rangle| = \sum_{i=1}^{n} w_i |\cos \theta_i|, \tag{2.8}$$

w — это вектор весов, который вычисляется с помощью матрицы сингулярных чисел Σ, подробное описание процедуры в [22]. Так как для кластеризации обычно строят матрицу расстояний (а не близости), поэтому из (2.8) получим

$$D_{\mathrm{Eros}}(V_x, V_y, w) = \sqrt{2 - 2 \sum_{i=1}^{n} w_i |\langle v_{x_i}, v_{y_i} \rangle|}. \tag{2.9}$$

Метод Борда широко применяется для определения победителя в играх голосования. Задачу многомерной кластеризации также можно представить как игру голосования, где кандидатами являются объекты. Таким образом, сначала МВР разбивается на n одномерных, строится матрица расстояний для каждой переменной, затем согласно правилу Борда определяются наиболее близкие к друг другу объекты, где «голосующими» являются размерности МВР. Авторы [23], [24] предложили использовать взвешенный метод Борда в силу того, что он учитывает на сколько сильно отличаются кандидаты. Обозначим кандидатов $s_0, \ldots, s_k$. Пусть s_0 — это кандидат, для которого ищется «сосед», тогда расстояние между ним и другими кандидатами $d_j = d_j(s_0, s_j), j \in \{1, \ldots, k\}$. Без потери общности считаем, что $d_{j-1} < d_j \ \forall j \in \{1, \ldots, k\}$. Вектор весовых коэффициентов w показывает значимость каждой переменной. Его можно задать, исходя из постановки задачи, или же вычислить, например так, как описано в [22]. Тогда количество очков, которое получит кандидат j вычисляется согласно

формуле

$$vs^j = \sum_{i=1}^{n} vs_i^j, \tag{2.10}$$

$$vs_i^j = w_i\big(1 + k(1 - \frac{d_j}{d_k})\big), \tag{2.11}$$

где $j = \overline{1,k}$ — номер кандидата; $i = \overline{1,n}$ — номер переменной; w_i — весовой коэффициент для переменной с номером i. Согласно полученным в (2.10) очкам определяется победитель, он и будет ближайшим к объекту s_0.

2.2. Кластеризация одномерных временных рядов

Имеем 3 набора данных, каждый из которых содержит информацию по показателю заболеваемости в районах Санкт-Петербурга с 1999 по 2014 гг. среди детей, подростков и взрослых. Построим матрицу расстояний для каждого случая, используя методы, представленные в предыдущем разделе в формулах (2.1), (2.2), (2.4). В кластерном анализе известны различные подходы: вероятностные, теоретико-графовые, логические, и др. Каждый из них имеет свои особенности, преимущества и недостатки. В связи с небольшим количеством объектов кластеризации (18 районов) в задаче удобно использовать иерархический подход, а именно построение дендрограммы. Это позволит визуализировать удаленность объектов друг от друга и принять решение о числе кластеров.

В качестве примера показаны дендрограммы, построенные по матрицам расстояний, рассчитанные согласно евклидовой метрике рис. 2.2 (а)–(в) и регулируемому индексу различия с коэффициентом $k = 1$ и расстоянием Евклида в качестве δ_{conv} (2.3)–(2.5). Красным цветом показаны кластеры; если обрезать дерево выше, то это приведет к уменьшению их количества, если ниже — к увеличению. Четыре кластера наиболее четко выделяются по всем возрастным группам и с применением различных метрик. Поэтому принято решение далее рассматривать анализ именно трех кластеров.

Можно заметить на рис. 2.1(в), что образовался кластер, содержащий всего один объект, что объясняется необычным скачком в показателе

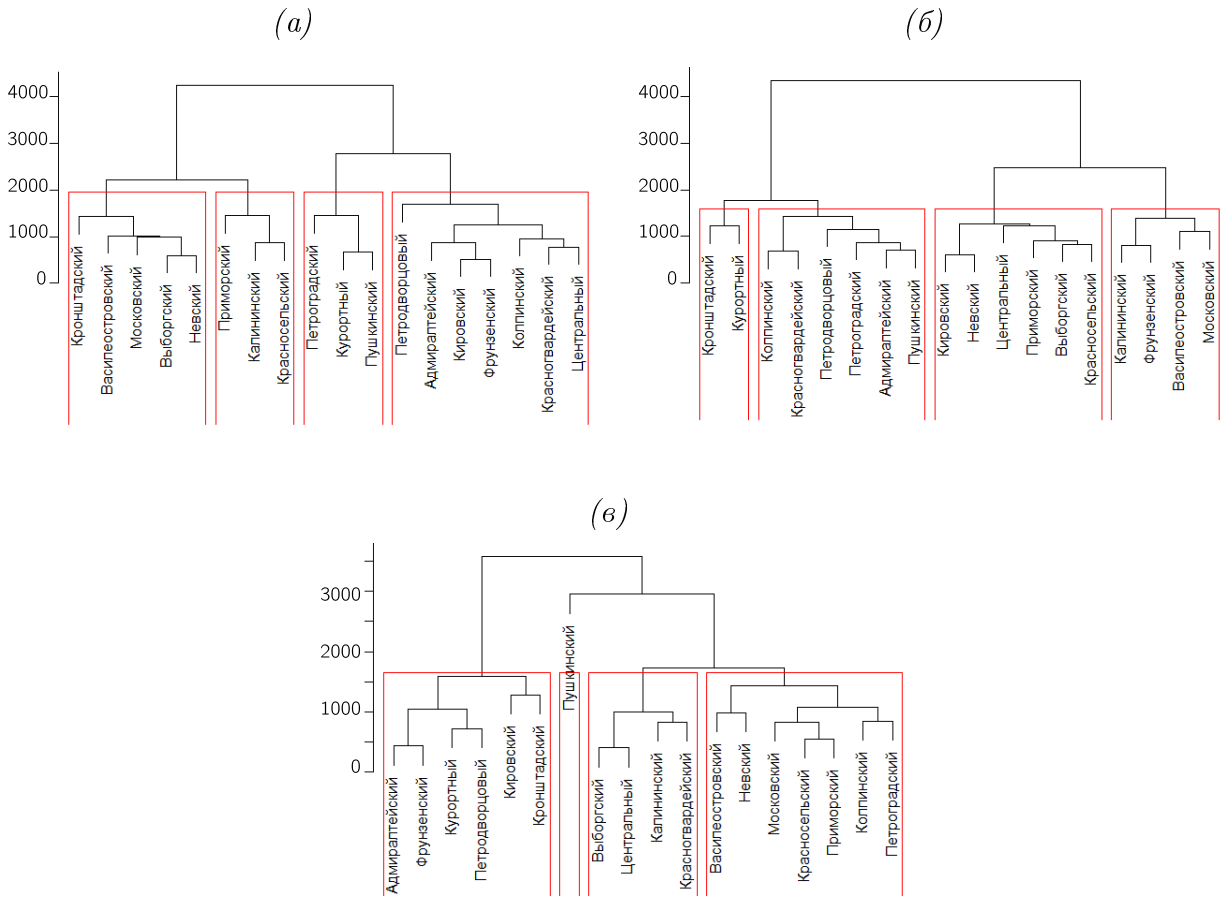


Рис. 2.1. Дендрограмма, построенная по матрице евклидовых расстояний между показателями заболеваемости районов:
(а) — детскими, *(б)* — подростковыми, *(в)* — взрослыми.

заболеваемости Пушкинского района рис. 2.3, это может быть связано с резким приростом населения района: по данным администрации увеличение трудоспособного населения за 2013–2015 гг. составило 60,9%.

В отличии от классификации данных кластеризация не имеет так называемого «правильного» решения. И результаты во многом зависят от выбора подхода, меры различия объектов и способа определения количества кластеров. Именно поэтому в задаче построено различных 4 модели для каждой возрастной группы, отличающиеся мерами: евклидова метрика, расстояние Фреше, регулируемый индекс различия с $k = 1$ и $k = 2$. Если считать, что к одной группе принадлежат объекты с похожими значениями наблюдений и динамикой, то чем ниже коэффициент вариации в кластере, тем качественнее произведено разбиение. Кроме того, можно найти центр кластера и построить вокруг него коридор.

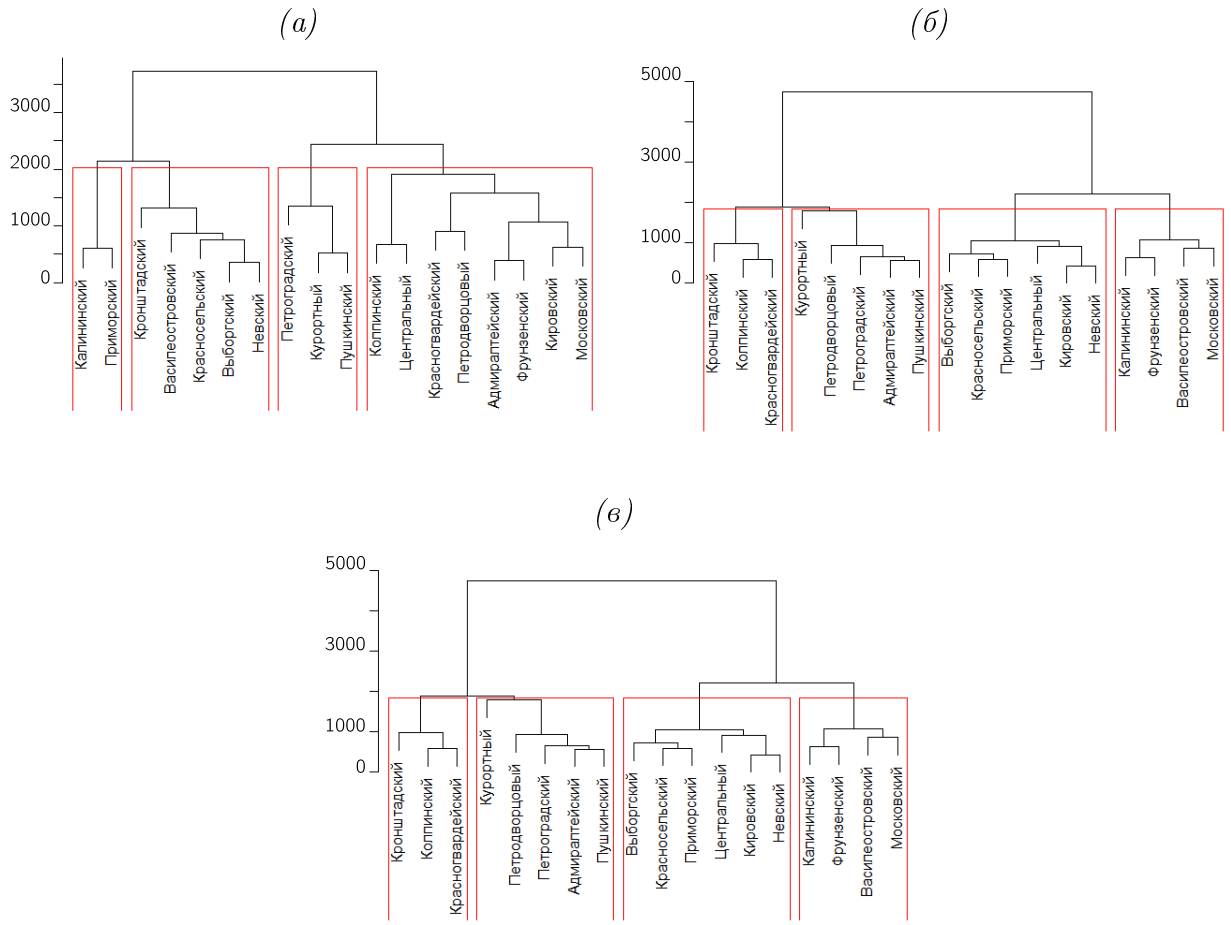


Рис. 2.2. Дендрограмма, соответствующая регулируемому индексу различия показателей заболеваемости районов:
(а) — детскими, *(б)* — подростковыми, *(в)* — взрослыми.

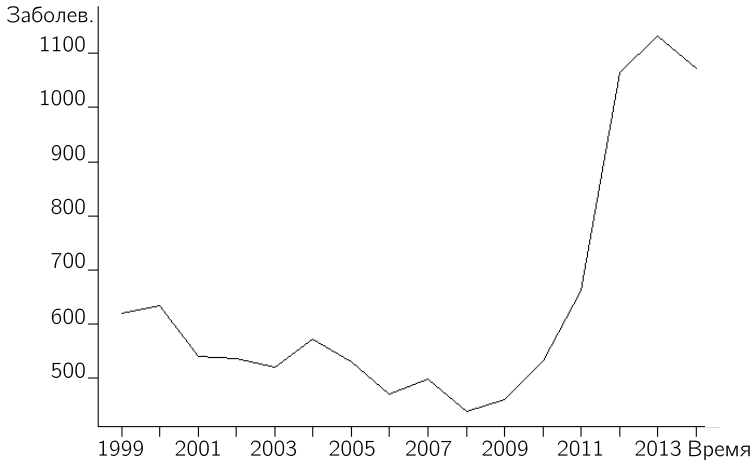


Рис. 2.3. Показатель заболеваемости взрослого населения Пушкинского района за 1999–2014 гг.

Большая площадь пересечения коридоров кластеров может быть признаком некорректного разбиения. На рис. 2.4 (а)–(б) представлены примеры коридоров двух похожих кластеров. Отличие представленных моде-

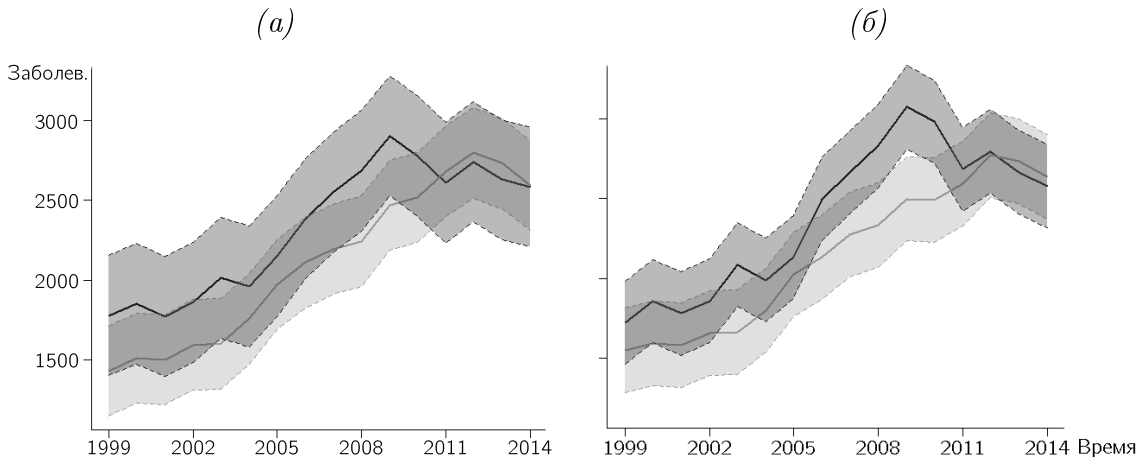


Рис. 2.4. Центры (—) и коридоры (- - -) двух кластеров, построенных с помощью евклидовой метрики *(а)* и регулируемого индекса различия *(б)* по показателю детской заболеваемости.

лей можно заметить при сравнении рис. 2.1(а) и рис. 2.2(а). В частности, Московский район отнесен к разным группам. В результате коридоры кластеров второй модели рис. 2.4(б) оказались уже. Отметим, что сначала в исследовании было построено не 4, а 6 моделей, но 2 модели, где мерой выступал регулируемый индекс сходства с $k = 1$ и $k = 2$ и расстоянием Фреше в качестве δ_{conv} (2.3)–(2.5), были исключены в силу полного перекрытия коридоров.

2.3. Кластеризация многомерных временных рядов

При многомерной кластеризации различают 2 подхода: сопоставление переменных (match-by-dimension) и предельное сопоставление (overall matching). Алгоритм кластеризации методами из первой группы следующий:

1) разбить многомерный ряд на несколько одномерных;
2) выполнить их кластеризацию;
3) агрегировать полученные результаты.

Методы второго подхода работают с целым временным рядом, чтобы не потерять важные связи между переменными. В основе часто лежит метод главных компонент или другие способы сокращения размерности. Для за-

дачи использовался метод Eros, однако, результаты оказались спорными: все районы оказались в одной группе. Это может быть связано с тем, что применение методов по сокращению размерности на небольших данных может сильно приуменьшать различия объектов.

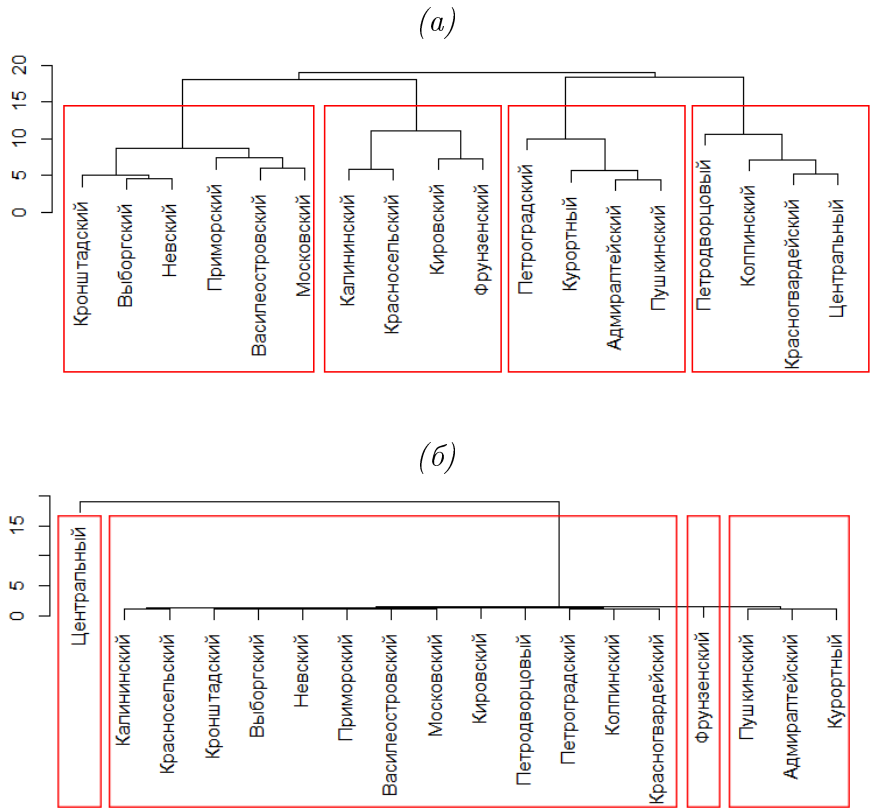


Рис. 2.5. Дендрограмма, соответствующая несимметричной матрице расстояний, полученной методом Борда, *(а)* построена по верхней треугольной матрице, *(б)* — по нижней; одномерная кластеризация проведена мерой Фреше.

В предыдущем разделе описан процесс одномерной кластеризации. Применим взвешенный метод Борда (описан в разделе 2.1) для агрегирования информации о гомологичных группах жителей Санкт-Петербурга различного возраста. Применяя этот метод, из трехмерной матрицы расстояний получим одномерную, по которой построим дедрограмму и выделим кластеры. Весовые коэффициенты вычислены согласно алгоритму в [22] и составляют 0.8803, 0.1028, 0.0169 для детской, подростковой и взрослой заболеваемости соответственно. Наибольшее влияние на результат оказала детская заболеваемость, что согласуется с фактом высокой значимости статистических показателей детей в здравоохранении.

Недостатком процедуры агрегирования информации, описанной в

разделе 2.1, является нарушения свойства симметричности, т.е. $dist(X,Y) \neq dist(Y,X)$. Это происходит из-за того, что для каждого района максимальное расстояние до других объектов различно, таким образом в (2.11) для каждого района d_k будет своим. Таким образом применяя этот подход получаем несимметричную матрицу расстояний. Сравним рис. 2.5(а) и рис. 2.5(б), которые представляют дендограммы построенные по одной и той же матрице. Очевидно, что результаты кардинально отличаются, что затрудняет интерпретацию.

Подход, представленный в (2.10)–(2.11) основан на переходе к безразмерным величинам, это позволяет избежать проблем, связанных с величинами разных порядков, которые могут присутствовать в многомерных рядах. Если модифицировать формулу, заменив максимум на глобальный, ассиметричности не будет (рис. 2.6). Дендрограммы для всех 4 моделей представлены в *Приложении 2*, расположение кластеров на карте Санкт-Петербурга — в *Приложении 3*.

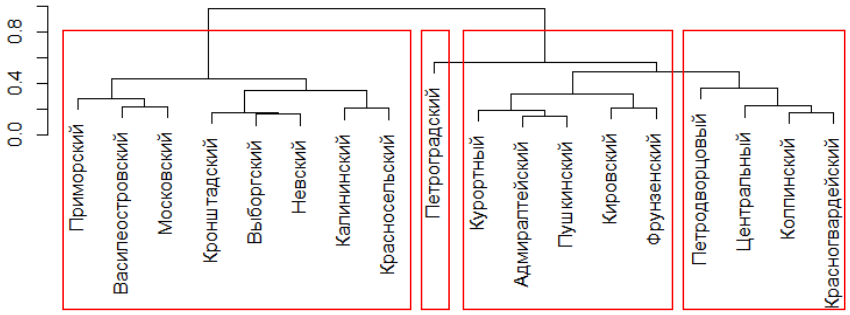


Рис. 2.6. Дендрограмма для проведения многомерной кластеризации сопоставлением переменных, одномерная кластеризация проведена расстоянием Фреше.

Можно заметить, что результаты сильно разнятся в зависимости от выбранного метода. Нельзя сказать, что одна из моделей является корректной, а другая —нет. Чтобы упростить интерпретацию результатов были найдены так называемые «стабильные» объекты, т.е. такие районы, которые находятся в одном кластере вне зависимости от метода. Для всех четырех построенных моделей оказалось всего две группы таких объектов: Калининский (4), Фрунзенский (17) и Красногвардейский (7), Невский (12). Можно отметить, что первая пара районов имеет примерно одинаковую численность и сформированы в один год, а вторая пара районов граничит территориально.

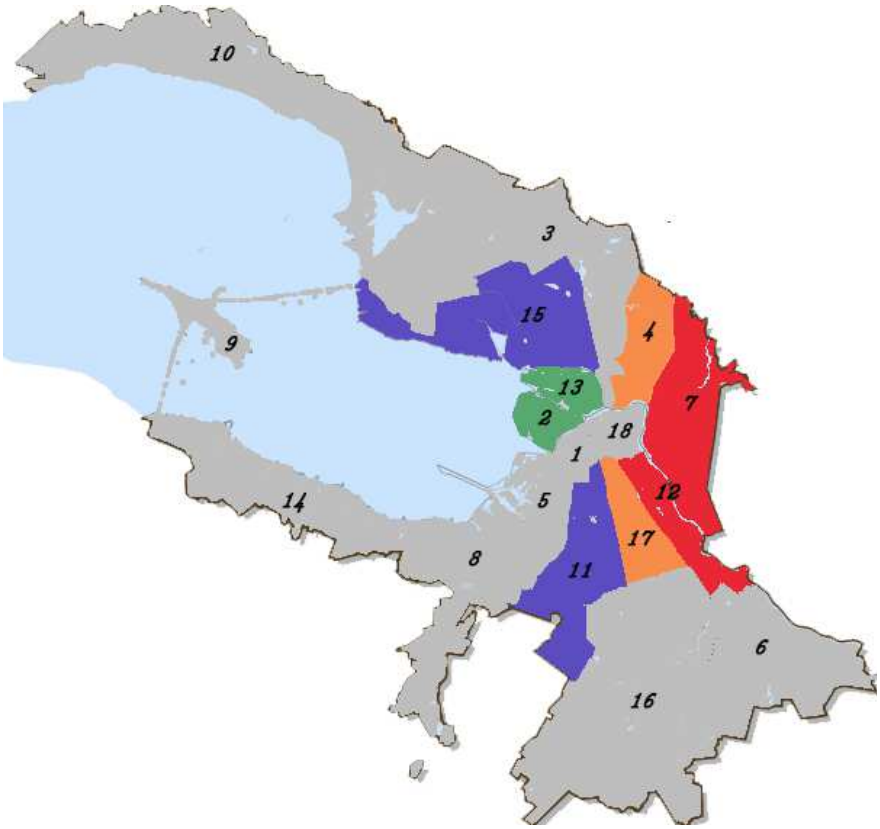


Рис. 2.7. Стабильные кластеры. Районы: *1* — Адмиралтейский, *2* — Василеостровский, *3* — Выборгский, *4* — Калининский, *5* — Кировский, *6* — Колпинский, *7* — Красногвардейский, *8* — Красносельский, *9* — Кронштадтский, *10* — Курортный, *11* — Московский, *12* — Невский, *13* — Петроградский, *14* — Петродворцовый, *15* — Приморский, *16* — Пушкинский, *17* — Фрунзенский, *18* — Центральный.

Если исключить из рассмотрения модель, построенную по евклидовой метрике, так как она не учитывает характер динамики, то получим 4 стабильных кластера рис. 2.7. Остальные районы меняли свое положение в зависимости от метода. Для того, чтобы получить окончательную картину можно использовать методы нечеткой кластеризации.

2.4. Выводы

Рассмотрена задача кластеризации районов Санкт-Петербурга по показателю заболеваемости. Сначала описан процесс кластеризации одномерных временных рядов, где из 6 различных моделей выбираются 4 наилучших. Критерием служит площадь пересечения коридоров, построенных вокруг центров кластеров. Далее эти результаты используются при проведении многомерной кластеризации подходом соответствия переменных.

Для агрегирования информации о трех возрастных группах исполь-

зуется взвешенный метод Борда. Однако, его применение оказалось под вопросом в связи с нарушением свойства симметричности полученных расстояний. Для решения этой проблемы произведена модификация метода и произведена многомерная кластеризация районов.

Метод Eros, который работает с многомерным временным рядом, как с единым объектом показал неоднозначные результаты, что, вероятно, связано с малым количеством данных. Получена карта стабильных кластеров. Вопрос распределения оставшихся районов по гомологичным группам рассмотрен в следующей главе.

Глава 3. Определение состава кластера в спорных ситуациях

В предыдущей главе выполнен кластерный анализ районов Санкт-Петербурга по показателю заболеваемости. Некоторые из них принадлежат одной группе вне зависимости от выбранного метода, такие объекты назовем стабильными. Относительно других нельзя сделать однозначные выводы. В связи с этим в задачах кластеризации используются подходы нечеткой логики. Каждому объекту соответствует степень принадлежности тому или иному кластеру.

Если в прикладной задаче известны наиболее важные критерии, которым должно соответствовать разбиение объектов, можно использовать эвристический алгоритм, предложенный в следующем разделе, для определения спорных объектов к кластерам.

В главе применение методов рассмотрено на примере анализа показателя детской заболеваемости в Санкт-Петербурге, так как возрастная группа «дети» имеет высокую значимость в здравоохранении. Более того, она оказывает наибольшее влияние на результат при многомерной кластеризации.

3.1. Критерии качества кластеризации

Сложность задачи кластеризации состоит в том, что непонятно какой результат является корректным. На него влияют и выбор подхода кластеризации, и выбор меры для построения матрицы расстояний, и определение количества кластеров. Поэтому существуют определенные критерии, по которым можно оценить качество кластеризации [25], [26]. Важно понимать, что здесь понятие «качество» имеет математический смысл. Со-

гласно одному из определений, кластеризация — это разбиение множества таким образом, что объекты из одной группы являются более схожими, чем объекты из разных групп. То есть, можно сказать, что кластеризация является «хорошей», если группы

- компактны — объекты располагаются близко друг к другу;
- отделимы — расстояние между объектами из разных групп значительно больше, чем расстояние между объектами одной группы;
- сконцентрированы вокруг центра, но это условие выполняется не всегда [14].

Условно критерии качества разделяют на

- внешние — используются, если есть дополнительная информация о структуре кластера;
- внутренние — основаны только на результатах кластеризации (близость объектов внутри группы, отделимость групп друг от друга и т.д.);
- относительные — применяются для сравнения две различных кластеризации.

Перед тем как рассмотреть некоторые критерии качества введем необходимые понятия. Будем считать, что данные представлены в виде матрицы $A = \{O_1, O_2, \ldots, O_N\}^T = \{V_1, V_2, \ldots, V_p\}$, где каждая строка $O_i,\ i = \overline{1, N}$ соответствует объекту, у которого зафиксировано p характеристик. Пусть $I_{\{k\}} = \{i \mid O_i \in C_{\{k\}}\}$ — множество индексов элементов, принадлежащих кластеру $C_{\{k\}},\ k = \overline{1, K}$, где K — количество кластеров. Тогда подматрица $A_{\{k\}}$ соответствует $C_{\{k\}}$ и состоит из строк матрицы A с индексами из $I_{\{k\}}$. Составим матрицу $X = \{v_1, v_2, \ldots, v_p\}$ из центрированных столбцов A: $v_j = V_j - \mu_j, j = \overline{1, p}$, где μ_j — среднее значение столбца j. Тогда матрица рассеяния $T = X^T X$, ее элементы представлены формулами:

$$t_{i,j} = \begin{cases} N\mathrm{Cov}(v_i, v_j),\ i \neq j, \\ N\mathrm{Var}(v_i),\ i = j, \end{cases} \tag{3.1}$$

где $i = \overline{1,p},\ j = \overline{1,p}$, T — симметрическая положительно полуопределенная матрица. Тогда суммарное отклонение (total sum of squares)

$$TSS = tr(T) = N \sum_{j=1}^{p} \mathrm{Var}(v_j), \tag{3.2}$$

геометрически TSS можно интерпретировать как величину рассеивания данных вокруг цетра масс. Аналогично можно определить матрицу рассеивания кластера $WG_{\{k\}} = X_{\{k\}}^T X_{\{k\}}$, элементы которой

$$w_{i,j} = \begin{cases} n_k \mathrm{Cov}(v_i, v_j),\ i \neq j, \\ n_k \mathrm{Var}(v_i),\ i = j, \end{cases} \tag{3.3}$$

где $i, j \in I_{\{k\}}, n_k = |C_{\{k\}}|$. Внутригрупповое отклонение (within-group sum of squares) опредеделяется как

$$WGSS_{\{k\}} = tr(WG_{\{k\}}), \tag{3.4}$$

общая внутригрупповая дисперсия:

$$WGSS = \sum_{k=0}^{K} WGSS_{\{k\}}. \tag{3.5}$$

Если обозначить центроид кластера $C_{\{k\}}$ как $\mu_{\{k\}}$, а центр масс объектов как μ. Определим матрицу $B = \{(\mu_{\{1\}} - \mu), (\mu_{\{2\}} - \mu), \ldots, (\mu_{\{K\}} - \mu)\}$, матрица рассеивания $BG = B^T B$, ее элементы выражены формулами

$$bg_{ij} = \sum_{k=1}^{K} n_k (\mu_{\{k\}i} - \mu_i)(\mu_{\{k\}j} - \mu_j). \tag{3.6}$$

Определим межгрупповое отклонение (between group dispersion)

$$BGSS = tr(BG) = \sum_{k=1}^{K} n_k \sum_{j=0}^{p} (\mu_{\{k\}j} - \mu_j)^2. \tag{3.7}$$

В таблице 3.1 приведены некоторые наиболее используемые индек-

Таблица 3.1. Индексы оценки качества кластеризации, не использующие дополнительную информацию о структуре кластера

№	Название или автор	Формулировка
1	Calinski-Harabasz	$\frac{N-K}{K-1}\frac{BGCC}{WGCC}$
2	Scott-Symons	$\frac{\det(T)}{\det(WG)}$
3	Scott-Symons (логарифмический)	$N\log\left(\frac{\det(T)}{\det(WG)}\right)$
4	Dann	$\frac{\min\limits_{k\neq l} d_{kl}}{\max\limits_{1\leq k\leq K} D_k}$
5	Marriot	$K^2\det(WG)$
6	Ray-Turi	$\frac{1}{N}\frac{WGSS}{\min\limits_{k<l}\|\mu_{\{k\}}-\mu_{\{l\}}\|^2}$
7	Scott-Symons	$\sum_{k=1}^{K} n_k \log\det\left(\frac{WG_{\{k\}}}{n_k}\right)$

сы. Индекс Калински-Харабасза (Calinski-Harabasz) предложен еще в 1974 году [27] и определен как произведение отношения взвешенной суммы квадратов расстояний между центроидами кластеров и центром масс (3.7) к сумме внутригрупповых отклонений (3.5) и нормирующего коэффициента.

Индекс Скота-Саймонса (Scott-Symons, № 2-3 в табл. 3.1) [28] — это отношение определителей суммарного отклонения (3.1) к сумме матриц рассеивания

$$WG = \sum_{k=0}^{K} WG_{\{k\}}. \tag{3.8}$$

Эти же авторы предложили логарифмический вариант индекса.

Индекс Данна (Dann) [29] показывает во сколько раз минимальное расстояние между объектами из разных кластеров больше максимального диаметра кластеров.

Можно заметить, что большинство индексов зависят от двух типов расстояний: внутригруппового и между элементами разных кластеров. Индекс Марриота (Marriot) [30] и Скотта-Саймонса (№ 7 в табл. 3.1) [28] напротив, учитывают только внутригрупповое рассеивание.

Индекс Рэя-Тури (Ray-Turi) [31] вычисляется как взвешенная сумма логарифмов от определителей матриц внутригрупповых отклонений. Этот

индекс не определен, если среди матриц $WG_{\{k\}}, k = \overline{1,K}$ есть нулевая.

Также часто применяют *валидацию по корреляции*, которая использует две матрицы: близости и инцидентности. Вторая строится следующим образом: каждому объекту соответствует строка и столбец, элемент матрицы равен 1, если объекты принадлежат одному кластеру. Чем выше значение коэффициента корреляции этих матриц, тем лучше кластеризация. Но этот метод не рекомендуется применять для моделей, основанных на плотности или смежности объектов.

Большое распространение получил индекс силуэта [32], так как для его вычисления нужны только расстояния между объектами. Пусть $D = \{d_{i,j}\}, i, j \in \{1, 2, \ldots, N\}$ — это матрица расстояний. Среднее расстояние от элемента кластера до других объектов этого же кластера обозначим

$$m_i = \frac{1}{n_k} \sum_{j \in I_{\{k\}}} d_{i,j},$$

где $i \in I_{\{k\}}$, а среднее расстояние до элементов другого кластера

$$m_{\{l\}}i = \frac{1}{n_l} \sum_{j \in I_{\{l\}}} d_{i,j}.$$

Если обозначить наименьшее из расстояний от объекта до других кластеров как

$$\underline{m_i} = \min_{l \in \{1,2,\ldots,N\} \setminus k} \{m_{\{l\}}i\},$$

тогда индекс силуэта объекта

$$s_i = \frac{\underline{m_i} - m_i}{\max\{\underline{m_i}, m_i\}}. \tag{3.9}$$

3.2. Построение нескольких кластеризаций районов по показателю детской заболеваемости

Кластеры будем искать алгоритмом $k-$медоид ($k-$medoid), он является более устойчивым по сравнению с алгоритмом $k-$средних, который также относится к группе вероятностных методов кластеризации. В

Таблица 3.2. Индекс силуэта и Данна для различных вариантов кластеризаций. Модели, в которых есть объекты, образующие самостоятельный кластер, выделены звездочкой.

Наименование меры различия	Индекс Данна			Индекс силуэта		
	2	3	4	2	3	4
FRECHET	0.49	0.33	0.39	0.45	0.34	0.23
EUCL	0.31	0.45	0.47	0.33	0.30	0.24
CORT(2)	0.16	0.17	0.17	0.32	0.32	0.23
CORT(3)	0.05	0.11	0.11*	0.25	0.33	0.29*
PER	1.28*	0.75*	0.94*	0.72*	0.58*	0.51*
ARMA	0.32*	0.32*	0.31*	−0.43*	−0.43*	−0.42*
CBD(0.4, 0.4)	0.54	0.31	0.49	0.32	0.17	0.22
CBD(0.8, 0.1)	0.37	0.50	0.49	0.32	0.30	0.26
CBD(0.6, 0.1)	0.32	0.42	0.50	0.31	0.23	0.26

отличии от иерархических методов они позволяют работать с большими объемами данных. Сначала задается количество кластеров k. После этого объекты распределяются по группам таким образом, чтобы центроиды кластеров находились на максимальном расстоянии друг от друга. Центроиды пересчитываются каждый раз при добавлении нового объекта. В случае подхода $k-$медоид центроид обязательно является объектом кластера, а вот в альтернативном подходе — это абстрактный объект (его называют центром масс), имеющий такую же природу, как у анализируемых данных. Для построения матрицы расстояния используем все метрики, представленные в главе 1. Из 21 возможного варианта отобрано 6 метрик: некоторые не прошли проверку на адекватность. Например, алгоритм динамической трансформации шкалы исключен в связи с тем, что в его основе лежит преобразование растяжения или сжатия и масштабирование — это является преимуществом при распознавании речи, но в задачах социального, экономического, демографического характера такие преобразования искажают реальные данные. Некоторые методы не подходят для работы с маленьким набором данных. В связи с тем, что для разбиения по группам нужно задать точное количество кластеров, будем строить несколько вариантов моделей для каждой меры различия.

Обзор критериев качества кластеризации приведен в предыдущем

разделе. Для задачи используем индекс силуэта (3.9) и индекс Данна (табл. 3.1), так как для их вычисления требуется только матрица расстояний и их интерпретация интуитивно понятна.

В табл. 3.2 рассчитаны значения индексов для различных моделей. Звездочкой отмечены индексы кластеризаций, где образовались кластеры, состоящие из одного объекта — в таком случае значение индекса нельзя принимать во внимание. Обратим внимание на то, что в различных кластеризациях такие объекты различны, т. е. нельзя говорить о том, что существует район, который значимо отличается от других. В связи с тем, что при использовании мер различия ARMA и PER все модели содержат такие объекты (составляющие самостоятельный кластер), они исключена из дальнейшего анализа. Адаптивный индекс различия (CORT) представлен в таблице двумя вариантами: с параметром равным 2 и 3. В среднем значения индексов качества лучше у модели с параметром 2, поэтому в дальнейшем будем рассматривать ее. По аналогичным соображениям оставим метрику CBD с параметрами 0.8 и 0.1.

Можно заметить, что в среднем индексы качества выше при распределении районов по трем кластерам. Таким образом, на данном этапе анализа имеем 4 варианта кластеризаций, в каждой выделено 3 группы относительно похожих районов Санкт-Петербурга по изменению показателя детской заболеваемости с 1999 по 2014 гг. Выделим стабильные кластеры (объекты, которые находятся в одном кластере вне зависимости от метода), для распределения остальных объектов далее предложен эвристический алгоритм.

3.3. Эвристический алгоритм распределения спорных объектов по кластерам

Как правило практические задачи требуют конкретного ответа на поставленные вопросы. Например, для решения нашей задачи нужно предоставить списки похожих районов. Полученные 4 модели не совпадают и это может вызвать затруднения в интерпретации у сотрудников организаций здравоохранения. То есть нужно прийти к одному вариан-

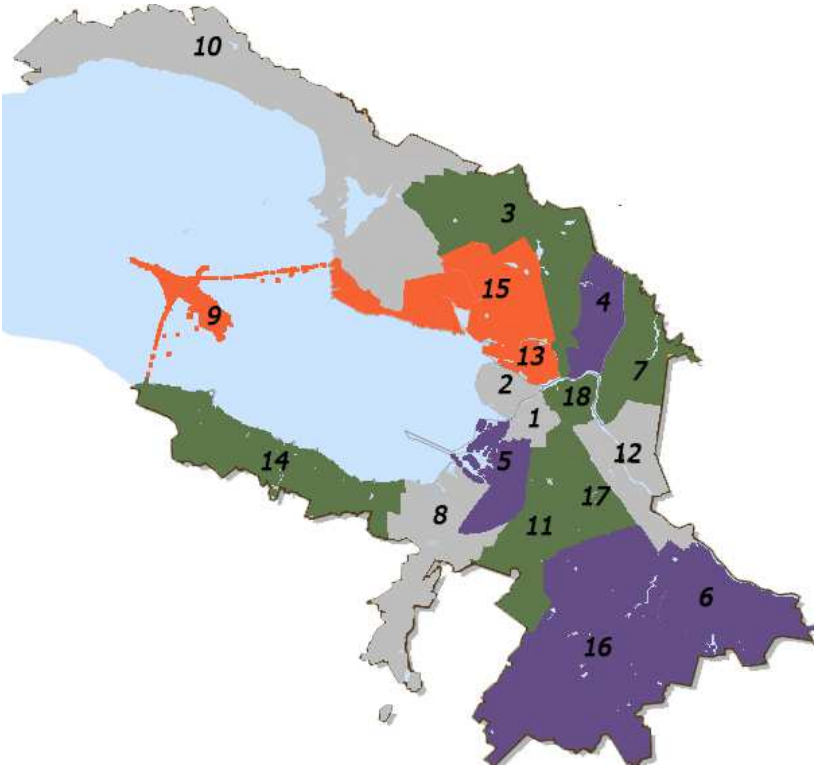


Рис. 3.1. Стабильные кластеры для кластеризации по детской заболеваемости. Районы: *1* — Адмиралтейский, *2* — Василеостровский, *3* — Выборгский, *4* — Калининский, *5* — Кировский, *6* — Колпинский, *7* — Красногвардейский, *8* — Красносельский, *9* — Кронштадтский, *10* — Курортный, *11* — Московский, *12* — Невский, *13* — Петроградский, *14* — Петродворцовый, *15* — Приморский, *16* — Пушкинский, *17* — Фрунзенский, *18* — Центральный.

ту распределения районов. Таким образом задача из области кластерного анализа сменяется задачей из области принятия решений.

В табл. 3.3 показано распределение районов по группам для каждой модели, полученной в предыдущем разделе. Можно заметить 3 стабильных кластера: {4, 5, 6, 16}, {3, 7, 11, 14, 17, 18}, {9, 13, 15}, оставшиеся 5 объектов {1, 2, 8, 10, 12} отнесены к различным группам в зависимости от меры различия.

Если известны критерии, характеризующие схожесть объектов, которые являются наиболее значимыми в определенной задаче, то можно воспользоваться итеративным эвристическим алгоритмом распределения оставшихся объектов по стабильным кластерам [33]. В алгоритме задача принятия решения о принадлежности объекта к кластеру рассматривается, как задача голосования [34], где избирателями выступают критерии, которым должны удовлетворять группы схожих объектов, а кандидатами — стабильные кластеры.

Таблица 3.3. Распределение районов по кластерам: *1* — Адмиралтейский, *2* — Василеостровский, *3* — Выборгский, *4* — Калининский, *5* — Кировский, *6* — Колпинский, *7* — Красногвардейский, *8* — Красносельский, *9* — Кронштадтский, *10* — Курортный, *11* — Московский, *12* — Невский, *13* — Петроградский, *14* — Петродворцовый, *15* — Приморский, *16* — Пушкинский, *17* — Фрунзенский, *18* — Центральный.

Наименование	Районы													
меры различия	1-2	3	4-6	7	8	9	10	11	12	13	14	15	16	17-18
FRECHET	c	b	a	b	b	c	b	b	c	c	b	c	a	b
EUCL	a	b	a	b	a	c	a	b	a	c	b	c	a	b
CORT(2)	a	b	a	b	b	c	a	b	a	c	b	c	a	b
CBD(0.8, 0.1)	a	b	a	b	b	c	b	b	a	c	b	c	a	b

Шаг 1. Поместим неопределенные объекты в список $L_{ind} = \{x_1, x_2, \ldots, x_l\}$.

Шаг 2. Зададим список $L_{cl}^k = \{c_1^k, c_2^k, \ldots, c_3^k\}$ кандидатов (потенциальных кластеров) для каждого неопределенного объекта $x_k \in L_{ind}$.

Шаг 3. Зададим ключевой критерий F_0, определяющий качество кластера. Предполагается, что с ростом (спадом) F_0 качество кластеризации возрастает.

Шаг 4. Зададим другие критерии $F_1, F_2, \ldots, F_m$, определяющие качество кластеризации, они являются избирателями в игре голосования. Как и на предыдущем шаге, предполагаем, что с ростом (спадом) $F_i, i \in 1, 2, \ldots, m$ растет качество кластеризации.

ИТЕРАЦИЯ

Шаг 5. Предположим, что объект $x_1 \in L_{cl}^k$ помещен в каждый кластер $c_i^1 \in L_{cl}^1$. Вычислим F_0 для вновь образованных кластеров $c_i^1 \in L_{cl}^1$. Этот шаг выполняется для каждого объекта из L_{ind}.

Шаг 6. Выберем кандидата с наибольшим (наименьшим) значением F_0, пусть это кластер c_d^n, образованный добавлением объекта x_n.

Шаг 7. Вычислим значения критериев $F_1, F_2, \ldots, F_m$ для каждого кандидата в списке L_{cl}^n.

Шаг 8. Воспользуемся методом Борда: упорядочим кандидатов в порядке убывания (возрастания) критерия F_i, первому кандидату из упорядочен-

ного списка назначаем m баллов, следующему — $(m-1)$ баллов и т.д. Этот шаг выполняется для каждого критерия $F_1, F_2, \ldots, F_m$.

Шаг 9. Найдем сумму баллов для каждого кандидата из списка L_{cl}^n. Кандидат с наибольшей суммой является победителем — c_e^n.

Шаг 10. Если кластер, полученный на предыдущем шаге c_e^n совпадает с кластером, определенным на шаге 6 — c_d^n, тогда переходим к следующему шагу, иначе возвращаемся к шагу 6, но исключаем c_d^n из игры на данной итерации.

Шаг 11. Включаем объект x_n в кластер c_e^n. Удаляем x_n из L_{ind}. Если список L_{ind} пустой, тогда задача распределения объектов по кластерам завершена, иначе возвращаемся к шагу 5 (новая итерация).

Как известно, применение метода Борда может привести к ситуации, когда несколько кандидатов имеют одинаковое количество очков, поэтому целесообразно использовать модификации метода.

Рассмотрим применение алгоритма на задаче.

Шаг 1. Поместим неопределенные объекты в список $L_{ind} = \{1, 2, 8, 10, 12\}$.

Шаг 2. Зададим список $L_{cl}^k = \{\{a, c\}, \{a, c\}, \{a, b\}, \{a, b\}, \{a, c\}\}$.

Шаг 3. Ключевой критерий F_0 — минимизация суммы абсолютных разностей сглаженных рядов, принадлежащих одному кластеру (сглаживание происходит скользящим средним с лагом равным 3).

Шаг 4. Другие критерии: F_1 — минимизация максимума абсолютных разностей последних 3 наблюдений; F_2 — минимизация максимального коэффициента вариации по наблюдениям; F_3 — минимизация отношения суммы абсолютных разностей последних трех наблюдений к этой же характеристике, но вычисленной для обновленного кластера (после добавления объекта); F_4 — минимизация отношения стандартного отклонения к этой же характеристике, но вычисленной для обновленного кластера.

ИТЕРАЦИЯ 1

Шаг 5. Представим, что объект $\{1\} \in L_{ind}$ принадлежит каждому потенциально возможному кластеру $\{a, c\}$ и вычислим F_0 для таких кластеров. Представим, что объект $\{2\} \in L_{ind}$ принадлежит каждому потенциально возможному кластеру $\{a, c\}$ и вычислим F_0 для таких кластеров. Эти дей-

ствия проделаем и для остальных объектов из L_{ind}.

Шаг 6. Минимум F_0 достигается, когда район 10 присоединяется к кластеру «b».

L_{ind}	1		2		8		10		12	
L_{cl}	a	c	a	c	a	b	a	b	a	c
F_0	1.76	1.83	1.58	1.66	1.76	1.55	2.06	1.50	1.88	1.53

Шаг 7. Вычислим значения критериев F_1, F_2, F_3, F_4 для кандидатов: кластера «a», состоящего из элементов $\{4, 5, 6, 16\} \cup 10$, и кластера «b» — $\{3, 7, 11, 14, 17, 18\} \cup 10$.

Кластер (кандидат)	Критерий			
	F_1	F_2	F_3	F_4
a	851.16	0.16	2.55	1.21
b	418.15	0.16	1.38	1.07

Шаг 8-9. Воспользуемся правилом Борда и преобразуем таблицу выше, найдем сумму баллов.

Кластер (кандидат)	Критерий (избиратель)				
	F_1	F_2	F_3	F_4	Сумма
a	1	2	1	1	5
b	2	2	2	2	8

Шаг 10. Победитель на шаге 9 и на шаге 6 совпадает, поэтому район 10 отнесен к кластеру «b».

Шаг 11. Преобразуем список неопределенных объектов $L_{ind} = \{1, 2, 8, 12\}$, удалив район 10, соответственно редактируем список $L_{cl}^k = \{\{a, c\}, \{a, c\}, \{a, b\}, \{a, c\}\}$. Так как L_{ind} не пустой, переходим к выполнению итерации 2.

После выполнения всех итераций алгоритма получим распределение всех районов по кластерам. Полученный результат зависит от выбранных критериев, то есть таких характеристик кластеризации, которые наиболее важны в конкретной прикладной задаче. Такой взгляд на задачу кластеризацию возможен, так как изначально отсутствует понятие «правильного» результата и нужно получить такое распределение, которое позволяет ответить на поставленные вопросы. В Санкт-Петербурге выделилось 3 группы

относительно однородных районов с точки зрения изменения показателя детской заболеваемости с 1999 по 2014 гг., их распределение по кластерам показано на рис. 3.2.

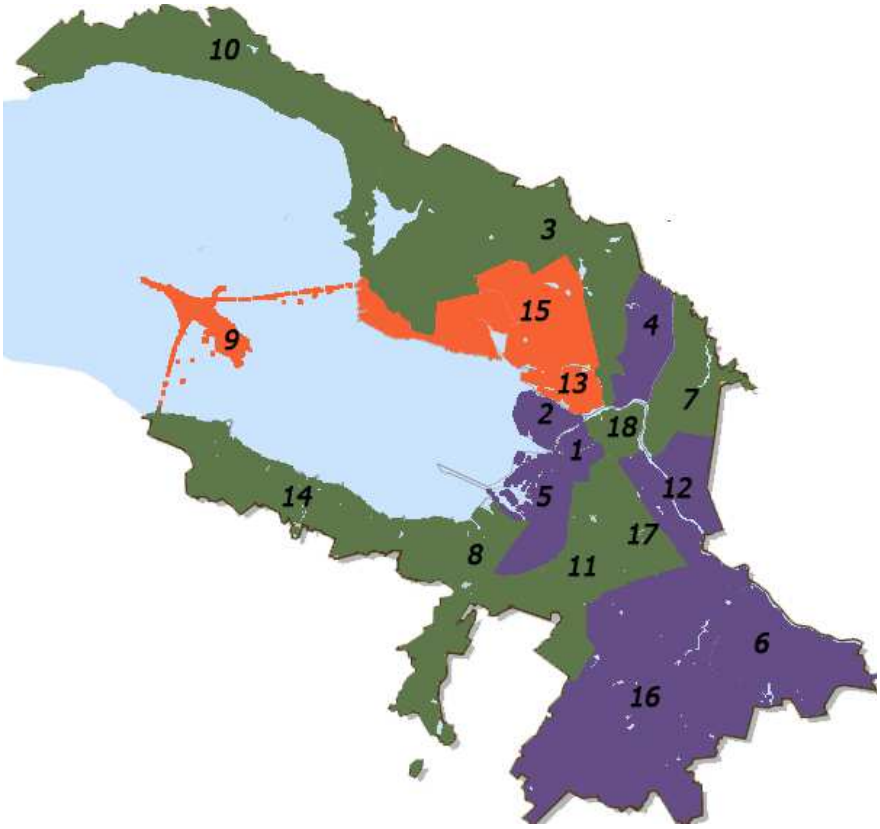


Рис. 3.2. Группы гомогенных районов по показателю детской заболеваемости. Районы: *1* — Адмиралтейский, *2* — Василеостровский, *3* — Выборгский, *4* — Калининский, *5* — Кировский, *6* — Колпинский, *7* — Красногвардейский, *8* — Красносельский, *9* — Кронштадтский, *10* — Курортный, *11* — Московский, *12* — Невский, *13* — Петроградский, *14* — Петродворцовый, *15* — Приморский, *16* — Пушкинский, *17* — Фрунзенский, *18* — Центральный.

3.4. Выводы

Особенность задачи кластеризации состоит в том, что нельзя определить является ли полученное распределение «правильным» в связи с отсутствием информации о структуре данных. Результат зависит от выбора подхода к распределению объектов по группам и меры различия. Определить наилучшую модель можно по индексам оценки качества кластеризации, но их, как и методов кластеризации, большое количество, что часто приводит к противоречию. В связи с этим в главе предложен новый эвристический метод, который позволяет определить объекты в один из

потенциально возможных кластеров. Выбор подходящей группы основан на критериях, которые формулируются, исходя из постановки задачи.

Заключение

Задача кластеризации временных рядов выделена отдельно в связи с тем, что применение классических метрик приводит к потере важных корреляций между наблюдениями. Чтобы определить являются ли временные ряды схожими, нужно учитывать не только их геометрическую отдаленность, но и характер динамики и изменчивости ряда. В области поиска нужных метрик проделана большая работа, описание наиболее используемых расстояний представлено в Главе 1. Эти методы показывают хорошие результаты при кластеризации больших данных, однако на коротких временных рядах они работают хуже. В связи с этим предложена новая метрика для кластеризации временных рядов, которая находит расстояние между характеристиками ряда (геометрическими, динамическими и характеристиками изменчивости). Эксперименты на искусственных и реальных данных показывают целесообразность использования алгоритма.

В следующем разделе представлено применение методов кластерного анализа временных рядов в решении прикладной задачи. Медицинским информационно - аналитическим центром предоставлены данные по заболеваемости жителей Санкт-Петербурга в возрастной разбивке: дети (до 14 лет), подростки (15–17 лет), взрослые (старше 18 лет) за период с 1999 по 2014 гг. Стоит вопрос являются ли районы города схожими относительно изменения показателя заболеваемости. В работе представлен анализ как одномерных временных рядов, так и многомерных. Задача многомерной кластеризации временных рядов является особенно сложной. Методы условно делятся на два подхода: алгоритмы первой группы учитывают корреляции между переменными, но являются сложными для интерпретации, большая часть из них основана на методе главных компонент; алгоритмы второй группы агрегируют информацию, которая получена при кластеризации каждой переменной отдельно.

Применение разных метрик и подходов кластеризации приводит к противоречивым результатам, поэтому в Главе 2 получена карта «стабильных» кластеров, то есть тех районов, которые оказались в одной группе при использовании различных методов. Однако существуют объекты, которые могут быть отнесены к нескольким кластерам. Для решения проблем такого рода обычно используют индексы, которые показывают на сколько ближе оказываются объекты, принадлежащие одному кластеру, относительно объектов, взятых из разных кластеров. Но при выборе модели по индексу качества кластеризации также можно столкнуться с противоречием, так как их большое количество. Поэтому в Главе 3 предложен новый эвристический метод, позволяющий однозначно распределить объекты по кластерам. Идея алгоритма состоит в том, что выбор нужного кластера — это игра голосования, где кандидатами являются различные варианты распределения объектов, а голосующими — критерии качества кластеризации, которые формулируются для каждой задачи отдельно.

Таким образом, в работе представлен обзор современных методов кластерного анализа временных рядов и предложены новые, описаны эксперименты на искусственных и реальных данных и рассмотрена прикладная задача по выявлению однородных групп районов
Санкт-Петербурга по уровню заболеваемости.

Литература

[1] Sankoff D., Kruskal J. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Ontario: Addison Wesley Publ. Company, 1983. 382 p.

[2] Berndt D. J., Clifford J. Using dynamic time warping to find patterns in time series // KDD workshop on knowledge discovery in databases. 1994. P. 359–370.

[3] Oates T., Firoiu L., Cohen P. R. Clustering time series with hidden Markov models and dynamic time warping // Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning. 1999. P. 17–21.

[4] Maharaj E. A. A significance test for classifying arma models // Journal of Statistical Computation and Simulation. 1996. Vol. 54, no. 4. P. 305–331.

[5] Corduas M., Piccolo D. Time series clustering and classification by the autoregressive metric // Computational Statistics & Data Analysis. 2008. Vol. 52, no. 4. P. 1860–1872.

[6] Fu T. C. A Review on Time Series Data Mining // Engineering Applications of Artificial Intelligence. 2011. Vol. 24, no. 1. P. 164–181.

[7] Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge, MA : The MIT Press, 2016. 775 p.

[8] Староверова К. Ю., Буре В. М. Мера различия временных рядов, основанная на их характеристиках // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2017. No 1. С. 51–60.

[9] Montero P., Vilar J. TSclust: An R package for time series clustering // Journal of Statistical Software. 2015. No. 62.1. P. 1–43.

[10] Montero P. M., Vilar J. A. Time Series Clustering Utilities. Feb. 2015, URL: https://cran.r-project.org/web/packages/TSclust/TSclust.pdf. (дата обращения 1.11.2016).

[11] Fan J., Zhang W. Generalised Likelihood Ratio Tests for Spectral Density // Biometrika. 2004. Vol. 91, no. 1. P. 195–209.

[12] Cilibrasi R., Vitànyi P. M. Clustering by compression // IEEE Transactions on Information Theory. 2005. Vol. 51, no. 4. P. 1523–1545.

[13] Alcock R. J., Manolopoulos Y. Time-Series Similarity Queries Employing a Feature-Based Approach // 7th Hellenic Conference on Informatics. 1999. P. 27–29.

[14] Сивоголовко Е. В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. 2011. No 4. С. 14–31.

[15] Keogh E., Leonardi S., Ratanamahatana C. A. Towards parameter-free data mining // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. P. 206–215.

[16] Chen Y., Keogh E., Hu B., Begum N., Bagnall A., Mueen A. and Batista G. The UCR Time Series Classification Archive. Jul. 2015, URL: http://www.cs.ucr.edu/eamonn/time_series_data/ (дата обращения 03.04.2017).

[17] Староверова К. Ю. Кластеризация временных рядов с использованием R // Процессы управления и устойчивость. 2016. Т. 3. No 1. С. 317–323.

[18] Буре В. М., Староверова К. Ю. Методы кластерного анализа как способ выявления неоднородности временных рядов на примере показателя заболеваемости в Санкт-Петербурге // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2016. No 4. С. 44–50.

[19] Староверова К. Ю., Буре В. М. Кластеризация районов Санкт-Петербурга по показателю заболеваемости // Расширенные тезисы IX Международной Петрозаводской конференции «Вероятностные методы в дискретной математике» (Петрозаводск, 30 мая – 3 июня 2016 г.). Петрозаводск, 2016. С. 92–94.

[20] Eiter T., Mannila H. Computing discrete Fréchet distance // Christian Doppler Laboratory for Expert Systems. 2014. 7 pages.

[21] Douzal Chouakria A., Nagabhushan P. N. Adaptive dissimilarity index for measuring time series proximity // Advances in Data Analysis and Classification. March 2007. Vol. 1, issue 1, pp. 1–43.

[22] Yang K., Shahabi C. A PCA-based similarity measure for multivariate time series // MMDB '04 Proceedings of the 2nd ACM Intern. workshop on Multimedia databases. 2004. P. 65–74.

[23] Li S. J. , Zhu Y. L., Zhang X. H., Wan D. BORDA counting method based similarity analysis of multivariate hydrological time series // Journal of Hydraulic Engineering. 2009. Vol. 40, no. 3. P. 378–384.

[24] Wang J. , Zhu Y. , Li S., Wan D., and Zhang P. Multivariate time series similarity searching // The Scientific World Journal. 2014. Article ID 851017. 8 pages.

[25] Milligan G. W. A Monte-Carlo study of 30 internal criterion measures for cluster-analysis // Psychometrica. 1981. Vol. 46, no. 2. P. 187–195.

[26] Desgraupes B. Clustering indices. Apr. 2013, URL: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. (дата обращения 07.02.2017).

[27] Calinski R. B., Harabasz J. A dendrite method for cluster analysis // Communications in Statistics. 1974. Vol. 3, no. 1. P. 1–27.

[28] Scott A. J., Symons M. J. Clustering methods based on likelihood ratio criteria // Biometrics. 1971. Vol. 27, no. 2. P. 387–397.

[29] Dunn J. Well separated clusters and optimal fuzzy partitions // Journal of Cybernetics. 1974. Vol. 4, issue 1. P. 95–104.

[30] Marriot F. H. B. Practical problems in a method of cluster analysis // Biometrics. 1975. Vol. 27, no. 3. P. 501–514.

[31] Ray S., Rose H. T. Determination of number of clusters in k-means clustering and application in colour image segmentation // Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques. 1999. P. 137–143.

[32] Peter J. R. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65.

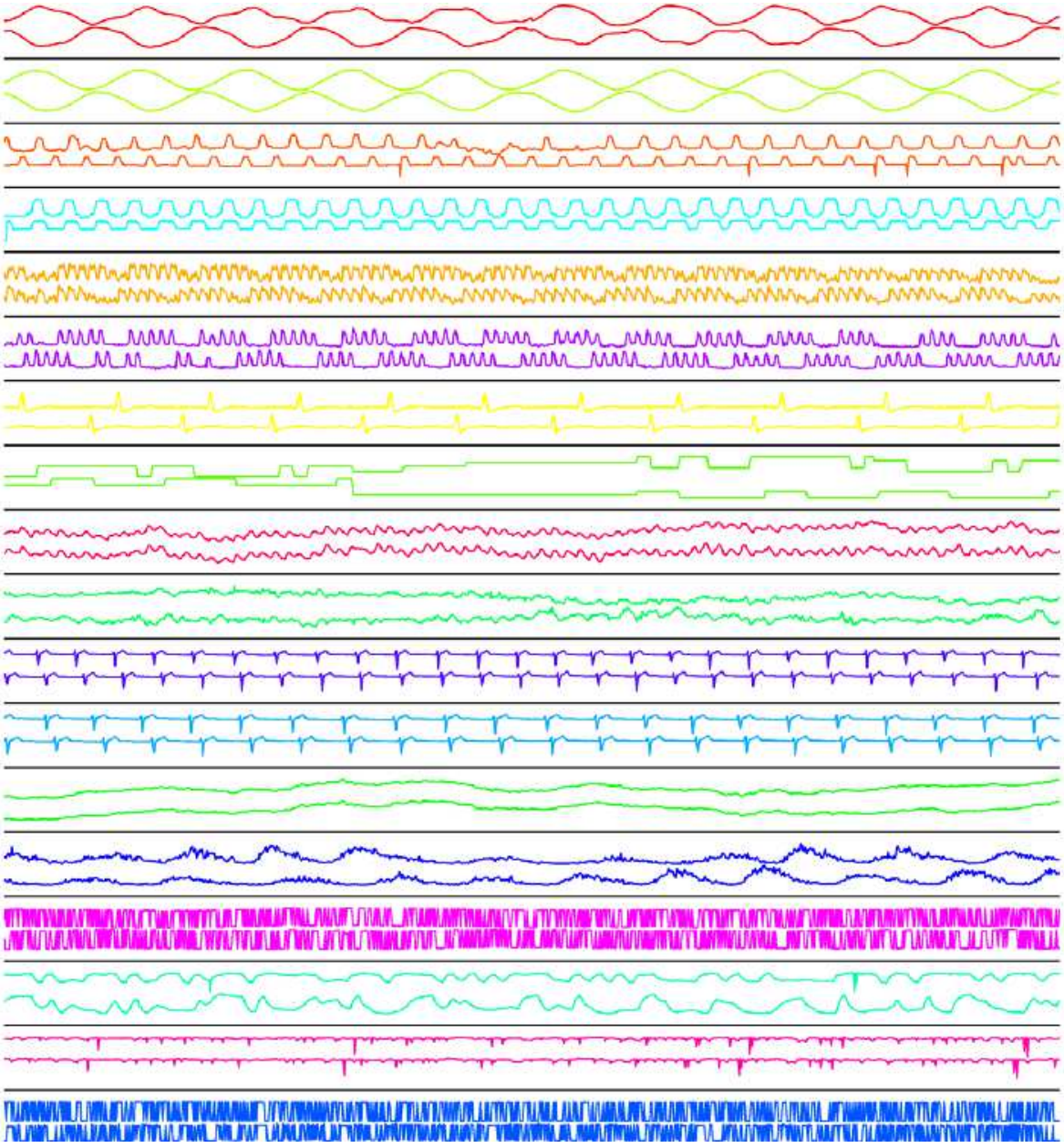
[33] Bure V. M., Staroverova K. U. Research on heterogeneity of children morbidity rate in Saint Petersburg // International Mathematical Forum. 2017. Vol. 12, no. 2. P. 77–85.

[34] Lippman D. Math in Society. 2.4: edit. Steilacoom: Pierce College. 2013. 428 p.

Приложения

Приложение 1

Набор временных рядов из различных предметных областей.

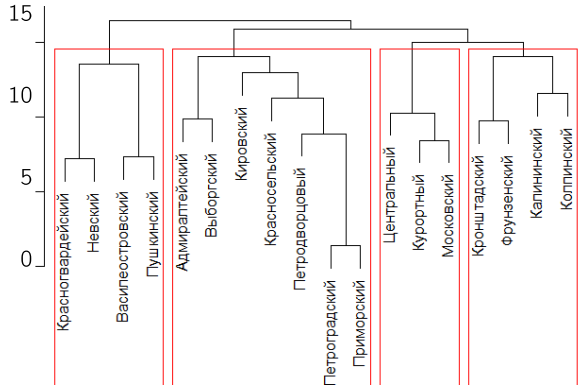


Приложение 2

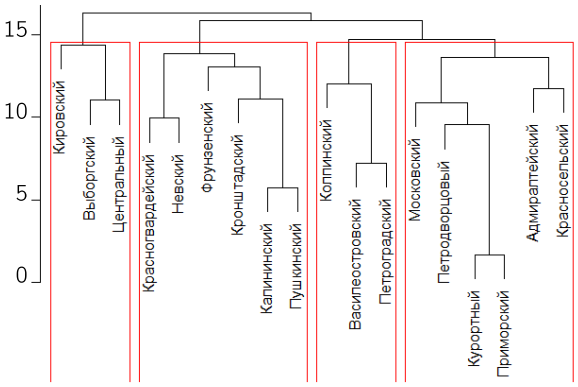
Дендрограммы. Кластеризация многомерных временных рядов.

Меры различия: *(а)* — расстояние Фреше, *(б)* — метрика Евклида, *(в)*-*(г)* — регулируемый индекс различия с $k = 1$ и $k = 2$.

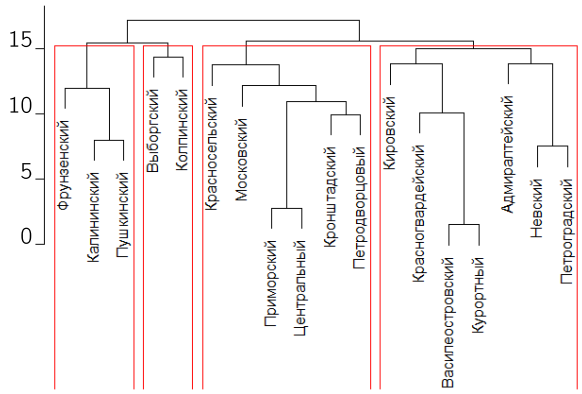
(а)



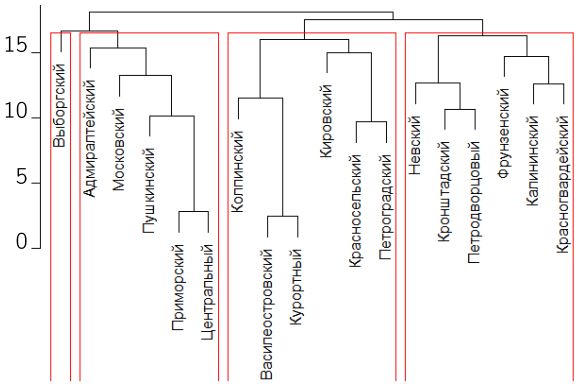
(б)



(в)



(г)



Приложение 3

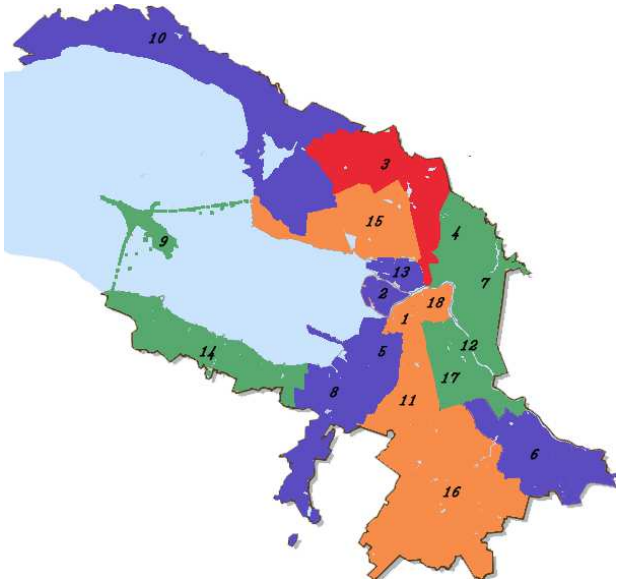
Результаты многомерной кластеризации.

Объекты, относящиеся к одной группе, имеют одинаковый цвет. Меры различия: *(а)* — расстояние Фреше, *(б)* — метрика Евклида, *(в)*-*(г)* — регулируемый индекс различия с $k = 1$ и $k = 2$. Районы: *1* — Адмиралтейский, *2* — Василеостровский, *3* — Выборгский, *4* — Калининский, *5* — Кировский, *6* — Колпинский, *7* — Красногвардейский, *8* — Красносельский, *9* — Кронштадтский, *10* — Курортный, *11* — Московский, *12* — Невский, *13* — Петроградский, *14* — Петродворцовый, *15* — Приморский, *16* — Пушкинский, *17* — Фрунзенский, *18* — Центральный.

(а)

(б)

(в)



(г)