УДК 338

А. С. Копырин
ФИЦ «Субтропический научный центр Российской академии наук», Сочи,
e-mail: kopyrin_a@mail.ru

КЛАСТЕРИЗАЦИЯ РЕГИОНОВ РОССИЙСКОЙ ФЕДЕРАЦИИ ПО СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМ ПОКАЗАТЕЛЯМ РАЗВИТИЯ ТУРИСТСКОЙ ОТРАСЛИ

Ключевые слова: регион, кластеризация, дерево решений, районирование, туризм.

В текущих изменяющихся условиях сегментация туристских территорий России может помочь определить потенциальные прорывные рынки. Своевременное выявление новых тенденций из огромного объёма данных играет важную роль в бизнес-процессах и принятии решений. Методы интеллектуального анализа данных помогают преодолеть этот дефицит знаний. Цель исследования – разделение регионов Российской Федерации на категории на основе имеющихся официальных статистических данных. Для достижения этой цели предполагалось решение двух взаимосвязанных задач. – Кластеризация регионов России на основе имеющихся статистических данных об их развитии в области туризма – Построение модели, интерпретирующей полученные результаты, для объяснения разбиения на кластеры. Подобная модель поможет сформировать основные цели развития для отстающих регионов для подтягивания их до уровня «локомотивов» отрасли. Указанные задачи решались с использованием методов интеллектуального анализа данных. Отдельные группы регионов, описанные в исследовании, могут помочь лицам принимающим решение лучше понять сферу туризма в регионе с точки зрения ее прибыльности и, соответственно, принять соответствующие стратегии по развитию регионов в зависимости от его текущего состояния. Дальнейшие исследования предполагают формирование общей модели цифровой трансформации сферы туризма в зависимости от специфики региона.

A. S. Kopyrin
FRC the Subtropical Scientific Centre of the Russian Academy of Sciences, Sochi,
e-mail: kopyrin_a@mail.ru

CLUSTERING OF THE REGIONS OF THE RUSSIAN FEDERATION ACCORDING TO SOCIO-ECONOMIC INDICATORS OF THE DEVELOPMENT OF THE TOURISM INDUSTRY

Keywords: region, clustering, decision tree, zoning, tourism.

In the current changing conditions, segmentation of tourist territories in Russia can help identify potential breakthrough markets. Timely identification of new trends from a huge amount of data plays an important role in business processes and decision-making. Data mining methods help to overcome this lack of knowledge. The purpose of the study is to divide the regions of the Russian Federation into categories based on the available official statistical data . To achieve this goal, two interrelated tasks were supposed to be solved. – Clustering of Russian regions based on available statistical data on their development in the field of tourism – Construction of a model interpreting the results obtained to explain the partitioning into clusters. Such a model will help to form the main development goals for lagging regions to pull them up to the level of the “locomotives” of the industry. These tasks were solved using data mining methods. The individual groups of regions described in the study can help decision makers better understand the tourism sector in the region from the point of view of its profitability and, accordingly, adopt appropriate strategies for the development of regions depending on its current state. Further research suggests the formation of a general model of digital transformation of the tourism sector, depending on the specifics of the region.

Введение

В текущих изменяющихся условиях сегментация туристских территорий России может помочь определить потенциальные прорывные рынки. Своевременное выявление новых тенденций из огромного объема данных играет важную роль в бизнес-процессах и принятии решений. Методы интеллектуального анализа данных помогают преодолеть этот дефицит знаний.

Учитывая растущее восприятие туризма как катализатора национального и регионального экономического развития, этот сектор становится предметом особого ин-

тереса ученых, представителей бизнеса и администрации, Правильное определение ключевых точек развития критически важно для привлечения туристов и инвесторов не только в «мейнстримовые» туристские регионы, но и в новые территории, которые имеют потенциал.

Для того, чтобы регион был соответствующим образом развит с точки зрения туризма, необходимо изучить общие социально-экономические условия для туризма. Исследование сосредоточено на разделении регионов Российской Федерации на категории на основе имеющихся официальных статистических данных [1, 2].

Для достижения этой цели предполагалось решение двух взаимосвязанных задач.

- Кластеризация регионов России на основе имеющихся статистических данных об их развитии в области туризма;

- Построение модели, интерпретирующей полученные результаты, для объяснения разбиения на кластеры. Подобная модель поможет сформировать основные цели развития для отстающих регионов для подтягивания их до уровня «локомотивов» отрасли.

Указанные задачи решались с использованием методов интеллектуального анализа данных.

Интеллектуальный анализ данных – это «процесс обнаружения знаний в базах данных» [3] для извлечения ранее неизвестных закономерностей в виде групп связанных записей данных (кластерный анализ), необычных наблюдений (обнаружение аномалий) и зависимостей (правило ассоциации) и т.п.

Кластерный анализ – широко используемый метод сегментации, который формирует группы случаев на основе предопределенных переменных, причем члены группы (кластера) должны иметь наиболее схожие переменные (принцип однородности), в то время как члены других групп непохожи (принцип неоднородности). Методы кластеризации минимизируют расстояние внутри кластера и максимизируют расстояние между кластерами для сегментации данных.

Для решения задачи построения модели использовался метод регрессионных решающих деревьев [4]. Построение дерева решений – популярный метод классификации данных различных классов (не менее двух классов). Этот метод в исходном формате не применим напрямую к кластеризации, поскольку наборы данных для кластеризации не имеют предварительно назначенных меток классов. Данная проблема была решена предварительной кластеризацией с помощью иерархического метода.

Материалы и методы исследования

Методика проводимой кластеризации

В исследовании используется иерархический кластерный анализ, чтобы найти туристические районы РФ, опираясь на социально-экономические критерии субъектов федерации. Термин «кластерный анализ» охватывает ряд различных алгоритмов и методов группировки объектов сходного типа в соответствующие категории. Это инструмент исследовательского анализа данных, целью которого является сортировка различных объектов по группам таким образом, чтобы степень ассоциации между двумя объектами была максимальной, если они принадлежат к одной группе, и минимальной в противном случае. В нашем случае метод группирует регионы страны в группы, где социально-экономические показатели развития туристского сектора максимально близки друг к другу и минимально близки к индикаторам в других группах. Согласно Кеттенрингу [5] именно иерархический кластерный анализ является наиболее широко используемой формой кластеризации, при решении практических задач

Алгоритмы агломерационной иерархической кластеризации можно охарактеризовать как жадные в алгоритмическом смысле. Последовательность необратимых шагов алгоритма используется для построения желаемой структуры данных. Предположим, что пара кластеров, объединяется или агломерируется на каждом шаге алгоритма. Тогда следуют эквивалентные представления одной и той же выходной структуры, построенные на n объектах: набор из n – 1 узлов, начиная с мелкого узла состоящего из n классов, и заканчивая тривиальным узлом, состоящим только из одного класса, все множество объектов; бинарное дерево (один или два дочерних узла в каждом неконечном узле), обычно называемое дендрограммой.

Традиционные методы кластеризации можно разделить на секционную кластеризацию и иерархическую кластеризацию [6, 7]. Секционная кластеризация определяет разбиение записей данных на k кластеров таким образом, что записи данных в кластере

находятся ближе друг к другу, чем записи в разных кластерах. Основная проблема методов данного типа заключается в том, что они часто очень чувствительны к исходным значениям и выбросам [8].

Иерархическая кластеризация представляет собой вложенную последовательность узлов. Кластеризация может быть создана путем построения дерева либо от листьев к корню (агломеративный подход), либо от корня вниз к листьям (разделительный подход). Подобных подход отличается от секционной кластеризации тем, что не группирует явно точки данных с использованием сравнения расстояний. Вместо этого выполняется кластеризация, путём классификации областей данных и пустых областей в пространстве.

С геометрической точки зрения дерево решений представляет собой разделение пространства данных. Последовательность разрезов от корневого узла до конечного узла представляет собой гиперпрямоугольник. Таким образом, что для числового атрибута алгоритм построения иерархического дерева выполняет бинарное разделение, т. е. каждый разрез разделяет текущее пространство на две части [9].

Алгоритм построения дерева при разделительном подходе включает две фазы: фазу роста и фазу обрезки. Фаза роста включает в себя рекурсивное разделение данных обучающего множества, приводящее к построению дерева иерархии, так что либо каждый конечный узел связан с одним классом, либо дальнейшее разделение данного листа приведет к тому, что по крайней мере его дочерние узлы окажутся ниже некоторого заданного порога. Этап сокращения направлен на обобщение полученной дендрограммы, созданного на этапе роста, путем создания поддерева, которое позволяет избежать чрезмерной подгонки к обучающим данным. Действия на этапе обрезки часто называют пост-обрезкой, в отличие от предварительной обрезки, которая происходит на этапе роста и направлена на предотвращение расщеплений, которые не соответствуют определенному заданному порогу.

Результаты кластерного анализа могут быть довольно субъективными, и зависеть от выбранного метода группировки объектов.

Наиболее часто используются следующие виды группировки:

1. Метод «полной связи» основан на максимальном (самом длинном) расстоянии объекта, состоящем из одной выборки из каждого кластера.

2. Метод «одиночной связи» – один из старейших методов кластеризации. Определяющим признаком этого метода является то, что расстояние между кластерами определяется как кратчайшая пара объектов с учетом только объектов, состоящих из одной выборки из каждого кластера.

3. Метод «средней связи» основан на схожем с предыдущими принципе. Однако расстояние между двумя кластерами является средним значением расстояний между всеми объектами выборки из каждого кластера.

4. Метод Уорда не вычисляет расстояние между кластерами, но объединение двух кластеров основано на величине суммы квадратов отклонений, чтобы максимизировать их внутреннюю однородность.

Рисунок 1 графически иллюстрирует эти четыре подхода.

В дальнейшем при исследовании будет использован метод Уорда, который предоставляет наибольшую внутреннюю связность полученных кластеров.

Результаты кластеризации будут представлены в виде двумерной диаграммы – дендрограммы.

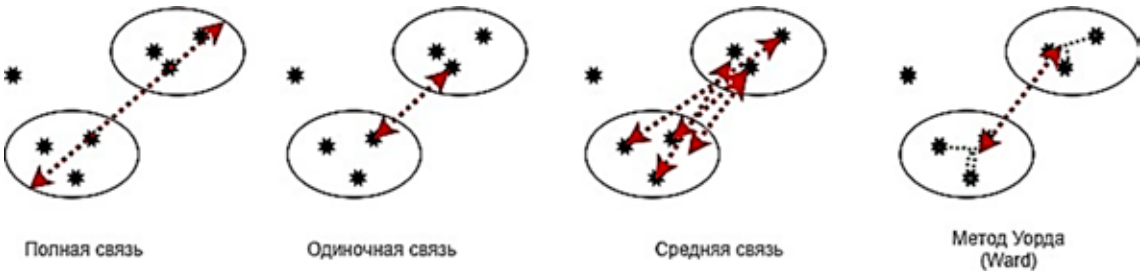

Рис. 1. Виды группировок при иерархической кластеризации

Листинг 1

```
# Вычисление расстояний
d <- dist(df_main_clean_n[,3:28], method = "euclidean")
# проведение иерархической классификации
hcd <- as.dendrogram(hclust(d, method = "ward.D2" ))
#визуализация полученного результата
nodePar <- list(lab.cex = 0.7, pch = c(NA, 19), cex = 0.7, col = "blue")
plot(hcd,  xlab = "Высота", nodePar = nodePar, horiz = TRUE,
     edgePar = list(col = 2:3, lwd = 2:1))
```

Методика построения модели

После кластеризации данных регионов, будет построена модель, объясняющая полученное разделение на группы (районы). Методом построения этой модели будет алгоритм дерева решений.

Одной из наиболее полезных характеристик деревьев решений является их понятность. Люди могут легко понять, почему дерево решений классифицирует экземпляр как принадлежащий к определенному классу. Поскольку дерево решений представляет собой иерархию тестов, неизвестное значение признака во время классификации обычно обрабатывается путем прохождения примера по всем ветвям узла, где было обнаружено неизвестное значение признака, и каждая ветвь выводит распределение классов. Результатом является комбинация различных распределений классов, сумма которых равна 1. В деревьях решений делается допущение, что экземпляры, принадлежащие к разным классам, имеют разные значения по крайней мере одного из своих признаков. Деревья решений, как правило, работают лучше при работе с дискретными/категориальными функциями.

При кластеризации и создании модели использовались данные Федеральной службы государственной статистики [2]. Обработка и визуализация данных проводилась с помощью языка R и интегрированной среды разработки Rstudio

Результаты исследования и их обсуждение

Кластеризация регионов

В качестве исходных данных использовались все доступные файлы статистики, полученные с официального сайта витрины данных федеральной службы статистики РФ, которые относятся к сфере туризма. Все файлы статистики были записаны в два набора данных – исходный и нормализованный. В нормализованном наборе данных, значения переменных нормализуются и шкалируются в разрезе каждой переменной. Затем данные обоих наборов данных очищаются, оставляя в них только данные о регионах.

Была проведена кластеризация по нормализованному набору данных. Для этого рассчитаем матрицу расстояний значимых переменных (без регионов и периодов времени). Расчёт выполним по формуле евклидова расстояния. По полученной матрице расстояний проведем иерархическую кластеризацию с использованием группировки метода Уорда и визуализируем результат (рис. 2).

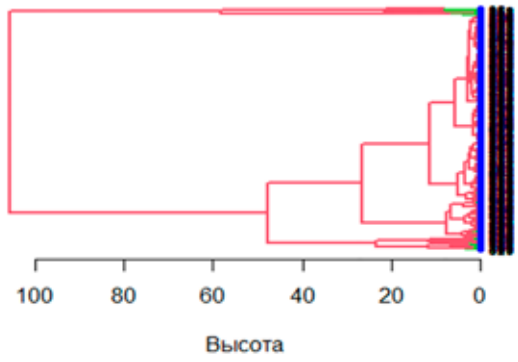

Рис. 2. Дендрограмма кластеризации регионов

Перейдём ко второй фазе кластеризации – обрезке полученного дерева. С помощью критерия оптимальности была рассчитана оптимальная высота полученного дерева – 15.

В результате получена тепловая карта (рис. 3) кластеров.

Прокомментируем полученные результаты. Исходя из перехода г. Москва из 3 в 7 кластер можно сделать вывод, что оптимальным количеством кластеров с учетом специфики задачи будет не 7, а 6. Также видно, что по некоторым регионам РФ (в частности Чукотскому АО и г. Севастополь) не хватает данных за некоторый период.

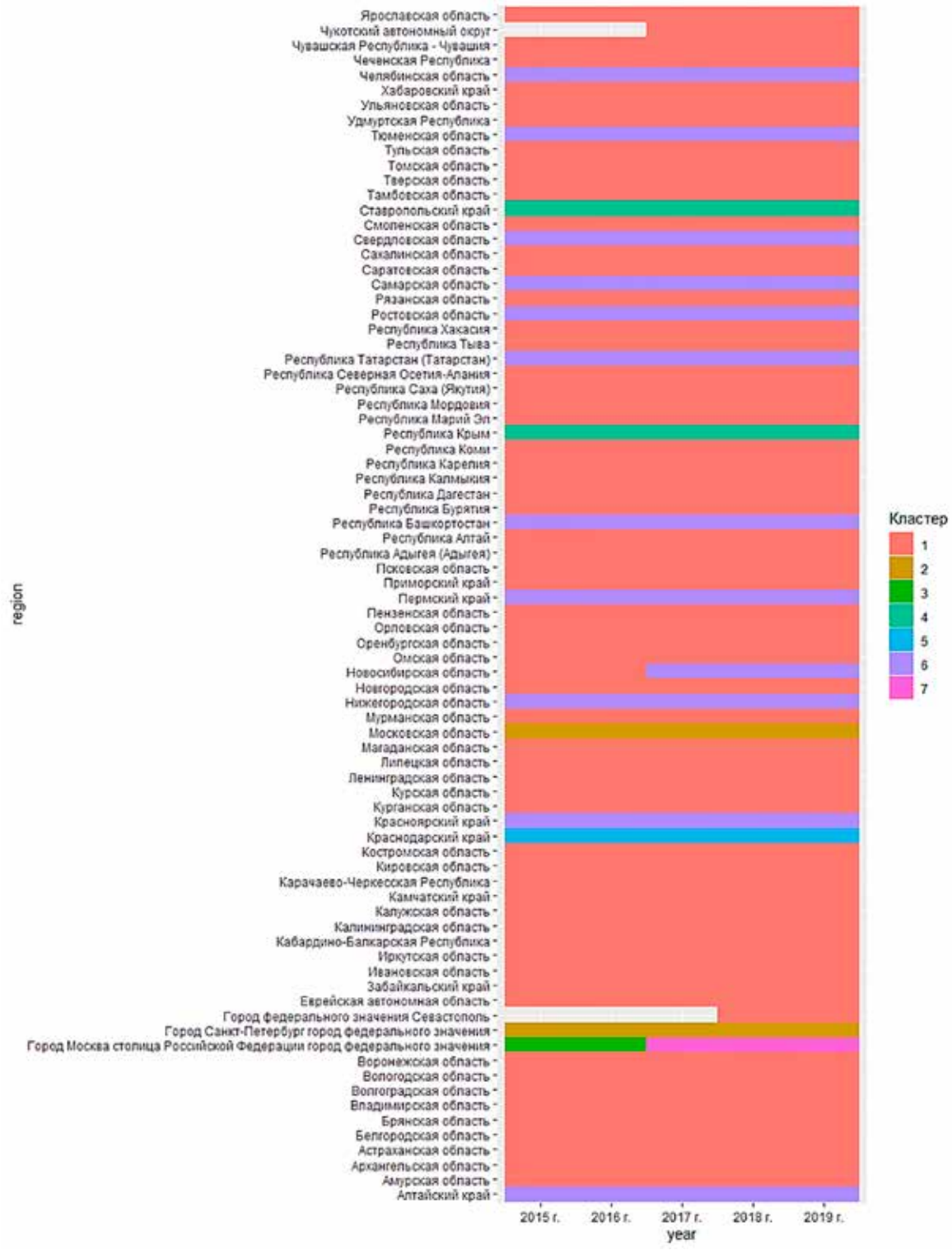

Рис. 3. Тепловая карта кластеризации регионов

Приведем описание полученных групп регионов (кластеров)

I. г. Москва (на рисунке кластеры 3 и 7), очевидно самый большой и платежеспособный регион.

II. Краснодарский край (на рисунке кластер 5) – регион с уникальной (для РФ) специализацией на санаторно-курортный туризм.

III. Московская область, г. Санкт-Петербург (на рисунке кластер 2) – регионы с большими

доходами от средств размещения, но с отсутствием значимой рекреационной базы

IV. Ставропольский край, Республика Крым (на рисунке кластер 4) – регионы с развитым туризмом, но с меньшими доходами и рекреационной базой по сравнению с предыдущими двумя кластерами

V. Алтайский край, Красноярский край, Нижегородская область, Новосибирская область, Пермский край, республика Башкортостан, Татарстан, Ростовская область, Самарская область, Свердловская область, Тюменская область, Челябинская область (на рисунке кластер 6) регионы со значимыми доходами средств размещения

VI. Все оставшиеся регионы (на рисунке кластер 1) – регионы с низким уровнем развития туристской отрасли

Следует отметить, что за период наблюдения Новосибирская область смогла перейти из одного кластера в другой, увеличив свои туристические доходы. После кластеризации построим модель, позволяющую отнести тот или иной регион к соответствующему кластеру.

Построение модели решающего дерева.

Разделим ненормализованный набор данных на обучающее и тестовое множество, построим дерево решений и визуализируем его.

Листинг 2.

```
set.seed(123)
inTrain <- createDataPartition(df_tree$cltr, p=0.9, list = F)
training <- df_tree[inTrain,]
testing <- df_tree[-inTrain,]
fit_tree <- rpart(cltr~., data=training[,3:29], method = 'class')
prp(fit_tree, type = 5, box.palette = "auto", varlen = 0, extra = 2,
    main = "Дерево решений кластеризации регионов")
```

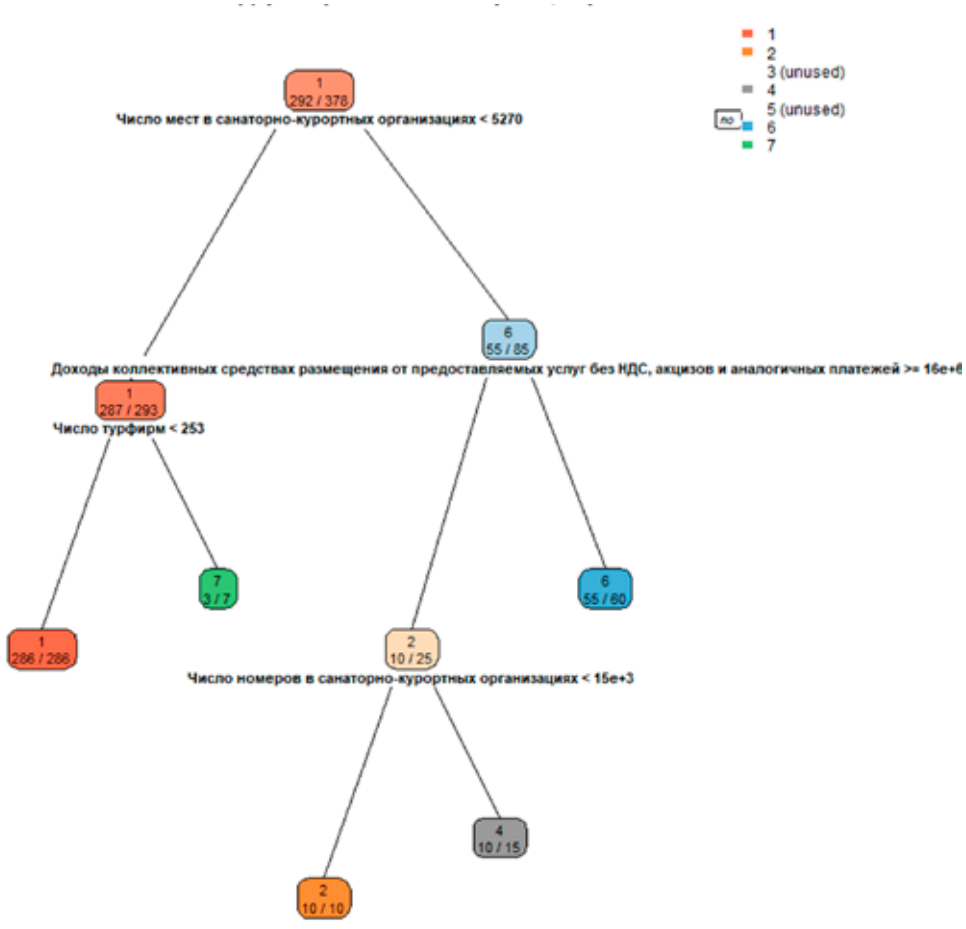

Рис. 4. Модель дерева решений кластеризации регионов

Из рисунка 4 видно, что модель не охватывает 2 построенных кластера 3 (Москва до 2017 года) и 5 (Краснодарский край), это обусловлено слишком малым количеством наблюдений в разрезе этих кластеров. Также на рисунке показано, сколько примеров из тестового множества были правильно распознаны в каждом кластере.

Как видно из информации, представленной на листинге 8 предсказательная сила модели на тестовом множестве очень велика. Точность модели составляет 100%, чувствительность и специфичность модели на распознанных классах также составляет 100%.

Главной задачей регионализации является четкое и системное представление предпосылок развития туризма. Эти предположения включают факторы со стороны предложения (местоположение и условия реализации) и факторы со стороны спроса (выборочные условия туризма) [10].

Использование современного стохастического метода позволяет категоризировать отдельные регионы и создавать кластеры, а также группы регионов, наиболее похожих друг на друга по выбранным входным переменным.

Применение кластерного анализа, процесс которого разбит на несколько последовательных шагов, показало значительную разницу в туристских регионах. В результате анализа туристские регионы разделены на шесть кластеров с точки зрения возможного развития и направления инвестиций в туризм. Указанное деление представлено в виде тепловой карты

Наиболее важными кластерами является регионы г. Москва и Краснодарский край. Кластером с наибольшим неиспользованным потенциалом являются регионы республики Крым и Ставропольского края.

Как уже упоминалось выше, регионализация туризма в Российской Федерации является одним из инструментов туристической политики на национальном уровне.

Построенная модель позволяет определить ключевые показатели, на которые необходимо обратить внимание при планировании административных программ развития туризма в регионе. К ним относятся: число турфирм, доходы коллективных средств размещения, число мест и номеров в санаторно-курортных учреждениях.

Заключение

В статье было представлено исследование, демонстрирующее возможность районирования регионов Российской Федерации, которое может быть выполнено с помощью методов интеллектуального анализа данных. Отдельные группы регионов, описанные в исследовании, могут помочь лицам принимающим решение лучше понять сферу туризма в регионе с точки зрения ее прибыльности и, соответственно, принять соответствующие стратегии по развитию регионов в зависимости от его текущего состояния.

В процессе интеллектуального анализа данных есть два этапа, которые очень важны и требуют больше всего времени: подготовка данных и интерпретация и оценка модели.

Дальнейшие исследования предполагают формирование общей модели цифровой трансформации сферы туризма в зависимости от специфики региона

Библиографический список

1. Витрина статистических данных. URL: https://showdata.gks.ru/finder/ (дата обращения: 11.09.2022).

2. Федеральная служба государственной статистики. URL: https://rosstat.gov.ru/ (дата обращения: 11.09.2022).

3. Murtagh F., Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min Knowl Discov. 2012. № 2. P. 86–97.

4. Liu B., Xia Y., Yu P. S. Clustering through decision tree construction. Proceedings of the ninth international conference on Information and knowledge management. 2000. С. 20-29.

5. Kettenring J.R. The practice of cluster analysis. Journal of classification. 2006. Т. 23. №. 1. С. 3-30.

6. Everitt B. Cluster analysis heinemann. 1974.

7. Jain A.K., Dubes R.C. Algorithms for clustering data. Prentice-Hall, Inc., 1988.

8. Bradley P.S. et al. Scaling Clustering Algorithms to Large Databases, Microsoft Research Report. 1998.

9. Quinlan J.R. Program for machine learning. 1993. С. 4-5.

10. Gáll J. Determining the Significance Level of Tourist Regions in the Slovak Republic by Cluster Analysis. Ekonomické rozhľady: vedecký časopis Ekonomickej univerzity v Bratislave.-Bratislava: Ekonomická univerzita v Bratislave. 2019. С. 451-462.