УДК 681.327.12.001.362

© 2021 г. **Н.С. Безруков**, канд. техн. наук,
Е.В. Полянская, канд. экон. наук
(Дальневосточный научный центр физиологии и патологии дыхания, Благовещенск)

СПОСОБ ПОСТРОЕНИЯ МОДЕЛИ КЛАСТЕРИЗАЦИИ ДАННЫХ НА ПРИМЕРЕ ДЕМОГРАФИЧЕСКИХ ПОКАЗАТЕЛЕЙ РЕГИОНОВ ДФО

Рассматривается задача построения модели классификации регионов ДФО по демографическим данным с помощью алгоритмов машинного обучения – метод стохастических вложений соседей с *t*-распределением, метод *K*-средних и самоорганизующиеся сети. Для демографических показателей построены столбчатые диаграммы и тепловые карты коэффициентов корреляции. Предложена замена демографических показателей на ранговые значения и рассмотрено влияние на результат классификации. На основе самоорганизующейся сети построен классификатор, позволяющий отнести регион ДФО к одному из классов: депрессивному, удовлетворительному или хорошему.

Ключевые слова: демографические показатели, метод стохастических вложений соседей с *t*-распределением, самоорганизующиеся сети, метод *K*-средних.

DOI: 10.22250/isu.2021.70.3-12

Введение

Кластеризация – объединение в группы схожих объектов – является одной из основных задач в области Data Mining [1]. Список прикладных областей, где она применяется велик: археология, медицина, психология, государственное управление, маркетинг, социология и другие дисциплины [2]. При анализе данных кластеризация часто бывает первым шагом. После выделения схожих групп применяются другие методы, для каждой группы строится отдельная модель. Например, в медицине можно разделить группы больных и затем для каждой группы применять отдельные тактики лечения.

В экономике можно для экономических субъектов (регионов или городов), находящихся в одной группе, применять единые программы развития и финансирования.

В период развития машинного обучения число методов кластеризации объектов довольно велико – несколько десятков алгоритмов и еще больше их модификаций [1, 3]. Однако часть их пришла из статистики и требует от разработчика соответствующих знаний.

Кластерный анализ предполагает решение следующих задач. Во-первых, необходимо отобрать данные и проверить их на ошибки. Затем применить какую-либо меру сходства между данными, качественно оценить данные, проверить их на вид распределения. Это позволит выбрать статистически подходящий метод кластеризации данных, а затем классифицировать. В завершение можно применить другой метод классификации – так, чтобы повторялось более 70% решений, что будет служить проверкой достоверности результатов [1, 3].

Вопросы демографических перекосов в разных субъектах Российской Федерации находят отражение в работах ведущих отечественных исследователей [4, 5]. Разработка социально-экономических программ для сглаживания таких перекосов является важной практической задачей, однако специалистам сложно разрабатывать программу для каждого региона в отдельности, а обобщенная программа для всех субъектов будет малоэффективной. Существует задача разделения регионов на классы [6, 7], в которых рационально использовать одни подходы для улучшения демографической ситуации и опускать другие. В работе предлагается построить модель кластеризации регионов на примере демографических показателей Дальневосточного федерального округа (ДФО).

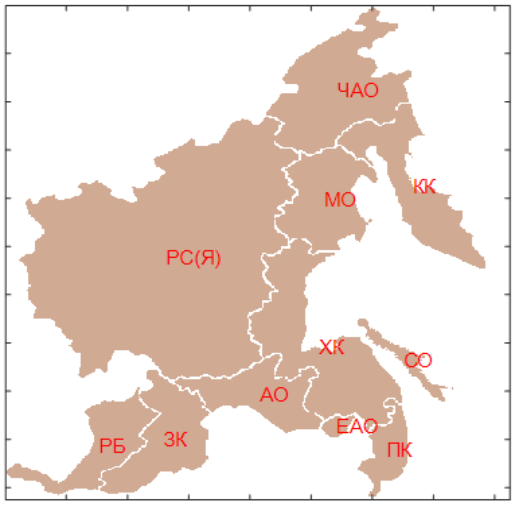

Рис. 1. Географическая карта ДФО.

Материал исследования

В качестве материала исследования использованы демографические статистические данные субъектов ДФО за период с 2000 г. по 2019 г. Географическая карта ДФО представлена на рис. 1 и состоит из 11 регионов: Амурская область (АО), Республика Бурятия(РБ), Еврейская автономная область (ЕАО), Забайкальский край (ЗК), Камчатский край (КК), Магаданская область (МО), Приморский край (ПК), Рес-

публика Саха (Якутия) РС(Я), Сахалинская область (СО), Хабаровский край (ХК), Чукотский автономный округ (ЧАО).

Основные статистические показатели демографии по регионам ДФО, на базе которых проводилось исследование, приведены в таблице, всего 20 показателей. Данные представлены в форме сводных таблиц за 20 лет – с 2000 г. по 2019 г., в которых имеется 4400 уникальных демографических показателей – 220 наблюдений, по 20 показателей в каждом наблюдении.

№ п/п	Наименование показателя
A1	Динамика численности постоянного населения (тыс. человек)
A2	Динамика среднего возраста населения
A3	Динамика удельного веса населения в трудоспособном возрасте
A4	Доля населения моложе трудоспособного возраста
A5	Доля населения старше трудоспособного возраста
A6	Динамика демографической нагрузки (на 1000 лиц трудоспособного возраста приходится лиц нетрудоспособных возрастов)
A7	Динамика демографической нагрузки (на 1000 лиц трудоспособного возраста приходится лиц моложе трудоспособного возраста)
A8	Динамика демографической нагрузки (на 1000 лиц трудоспособного возраста приходится лиц старше трудоспособного возраста)
A9	Динамика коэффициентов рождаемости
A10	Смертность населения в трудоспособном возрасте
A11	Динамика коэффициентов смертности населения (на 1000 населения)
A12	Динамика коэффициентов смертности населения в трудоспособном возрасте от всех причин (число умерших на 100 000 населения)
A13	Динамика коэффициентов смертности мужчин в трудоспособном возрасте от всех причин (число умерших на 100 000 населения)
A14	Динамика коэффициентов смертности женщин в трудоспособном возрасте от всех причин (число умерших на 100 000 населения)
A15	Динамика коэффициентов младенческой смертности (число умерших в возрасте до года на 1000 родившихся живыми)
A16	Динамика коэффициентов естественного прироста (убыли) населения (на 1000 жителей)
A17	Динамика ожидаемой продолжительности жизни при рождении
A18	Динамика коэффициентов миграционного прироста (убыли) населения
A19	Коэффициенты миграционного прироста (на 10 000 человек населения)
A20	Динамика коэффициентов общего прироста (убыли) населения (на 1000 населения)

Данные проверялись на корректность и ошибки. В единичных случаях данные отсутствовали и их пришлось заполнить с помощью моделей аппроксимации. В некоторых случаях данные были некорректно введены (лишний ноль и отсутствие запятой).

Качественная оценка данных

Предварительный статистический анализ каждого показателя производился с помощью столбчатых диаграмм в логарифмическом масштабе

(рис. 2), что обеспечило визуализацию сводной статистики для выборочных данных.

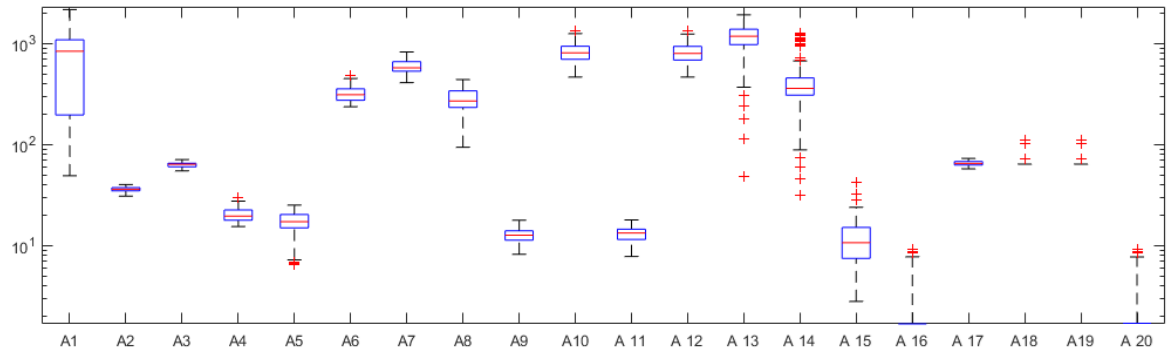

Рис.2. Столбчатые диаграммы демографических показателей.

Столбчатые диаграммы интерпретируются следующим образом:

1) нижняя и верхняя часть каждого прямоугольника – это 25 и 75 процентили выборки. Расстояние между низом и верхом каждого прямоугольника – межквартильный размах;

2) красная линия в середине каждого прямоугольника – медиана выборки;

3) усы – это линии, проходящие над и под каждым прямоугольником. Усы идут от конца межквартильного диапазона до самого дальнего наблюдения в пределах длины усов;

4) наблюдения за пределами длины усов отмечены как выбросы (значение, которое более чем в 1,5 раза превышает межквартильный размах). Выброс отображается красным знаком «+».

По построенной статистике данных можно сделать ряд выводов:

диапазон изменения по показателям различается на три порядка (как, например, А1 и А20);

динамика коэффициентов смертности мужчин в трудоспособном возрасте достоверно больше, чем у женщин, поскольку прямоугольники у показателей А13 и А14 не пересекаются;

динамика демографической нагрузки от молодежи достоверно больше, чем от людей пенсионного возраста, поскольку прямоугольники у показателей А7 и А8 не пересекаются;

у половины показателей параметры имеют выброс за полуторный межквартильный размах, т.е. эти параметры сильно меняются во времени (с 2000 г. по 2019 г.).

Анализ зависимости между показателями производился с помощью коэффициента корреляции. Значения коэффициента корреляции представлены на тепловой карте (рис. 3). Видно, что признаки А6-А8 имеют высокие значения коэффициента корреляции между собой, поскольку признаки ха-

рактеризуют одинаковую динамику демографической нагрузки только для различных возрастных групп. Такая же зависимость наблюдается в признаках A10-A12, характеризующих смертность населения. Многие показатели имеют корреляцию с годом наблюдения. Темно-синие области говорят, что показатели имеют между собой недостоверную корреляцию, таких зависимостей 88.

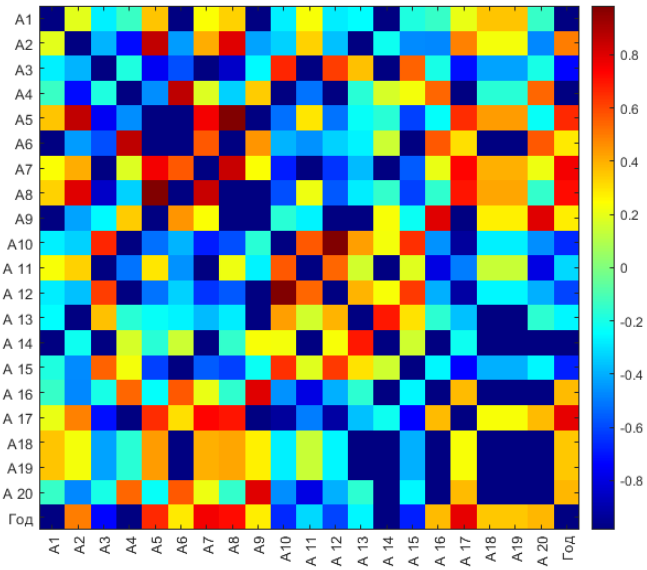

Рис. 3. Тепловая карта коэффициентов корреляции демографических показателей регионов ДФО.

Однако если рассмотреть зависимость между показателями более подробно для отдельного региона, то станет ясно, что эта корреляция наложена фактором времени. Например, как у показателей A3 и A10 для РБ. На рис. 4 видно, что оба показателя имеют линейную зависимость друг от друга (рис. 4а), которая обусловлена временной зависимостью (рис. 4б) – наличием фазы роста до 2005 г. и фазой падения с 2005 г. На данные наложен восходящий и нисходящий тренд во времени.

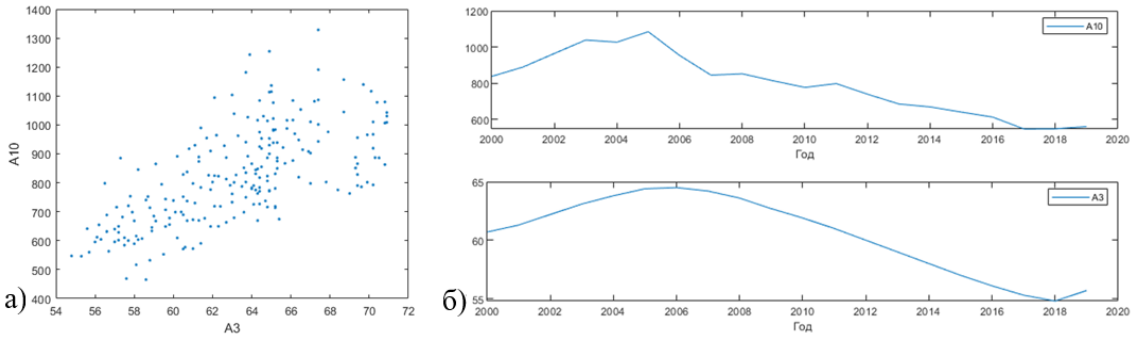

Рис. 4. Динамика показателей A3 и A10 для регионов ДФО (а) и зависимость от времени для РБ (б).

Из-за наличия больших разбросов в диапазонах и влияния временного фактора в виде трендов было предложено заменить реальные значения на ранговые по регионам по аналогии с работой [7]. Регионам в определенный

год по каждому показателю присвоили ранг от 1 до 11. Если реальное значение имело наименьшее значение ставили, 1, если наибольшее, то 11.

Тогда для ранговых данных тепловая карта коэффициентов корреляции примет вид, как на рис. 5. Видно, что признаки А6-А8, как и А10-А12, остались зависимыми. После преобразования недостоверных данных стало 166 (темно-синие квадраты), или почти в два раза больше, чем на рис. 3. Также и все корреляции с годом стали недостоверными.

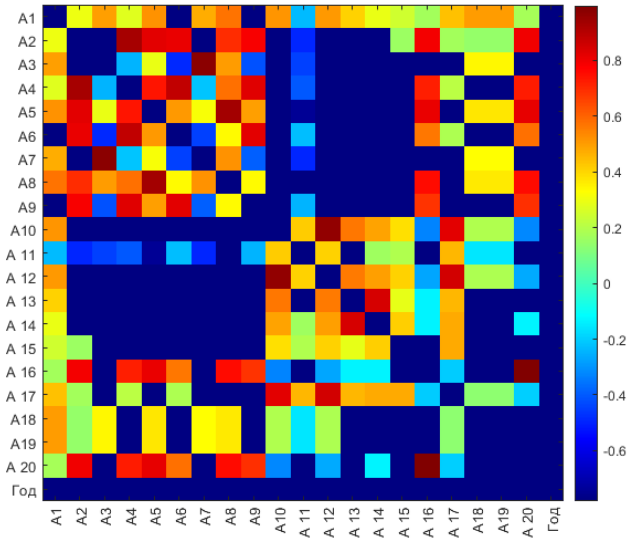

Рис. 5. Тепловая карта коэффициентов корреляции ранговых демографических показателей регионов ДФО.

Зависимость между показателями А3 и А10, рассмотренная на рис. 4, стала недостоверной. Зависимость между показателями А5 и А8 для региона РБ подтверждает, что корреляция от фактора времени отсутствует. На рис. 6а видно, что оба показателя имеют линейную зависимость, тогда как временная зависимость (рис. 6б) отсутствует, данные не имеют одинакового движения по годам.

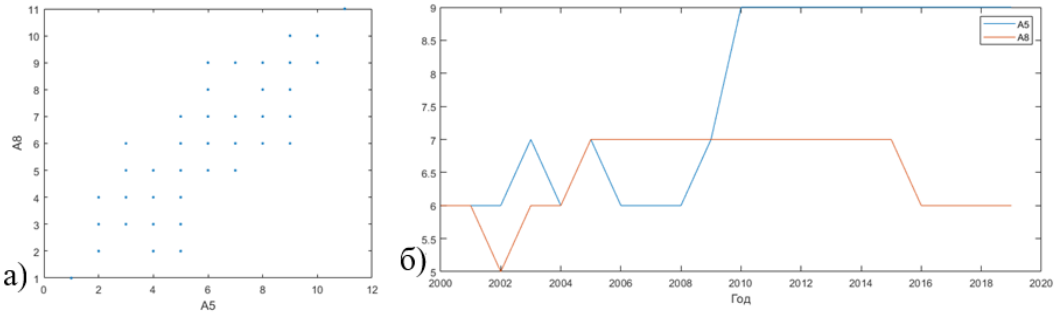

Рис.6. Динамика ранговых показателей А5 и А8 для регионов ДФО (а) и зависимость от времени для РБ (б).

Для предварительной оценки возможности разделения данных на классы воспользуемся методом стохастических вложений соседей с *t*-распределением (*t*-distributed Stochastic Neighbor Embedding, tSNE). Данный алгоритм машинного обучения использует технику нелинейного снижения

размерности. Алгоритм преобразует каждый объект наблюдения высокой размерности в объект из двух или трех показателей, которые затем легко представить на графике. Причем похожие объекты заменяются на близко расположенные точки на графике, а непохожие – на точки, расположенные далеко друг от друга.

Использование метода tSNE для реальных показателей подтвердило влияние времени на результат снижения размерностей (рис. 7а). Данные для большинства регионов вытянуты в линию. Анализ точек в каждой линии показал, что если в начале линии находится результат 2000 г., то в конце линии результат 2019 г. Ряд регионов (КК, МО, ЧАО, ЕАО) наложилися друг на друга. Если соседство регионов КК, МО, ЧАО можно объяснить тем, что они являются северными, с одинаковыми чертами, то наличие в этой группе ЕАО объяснить сложно.

Использование метода tSNE для ранговых показателей позволило сгруппировать регионы вокруг центров вне зависимости от года наблюдения (рис. 7б). Уже по этим данным профильный специалист может выделить группы и найти дискриминантные уравнения для классификации.

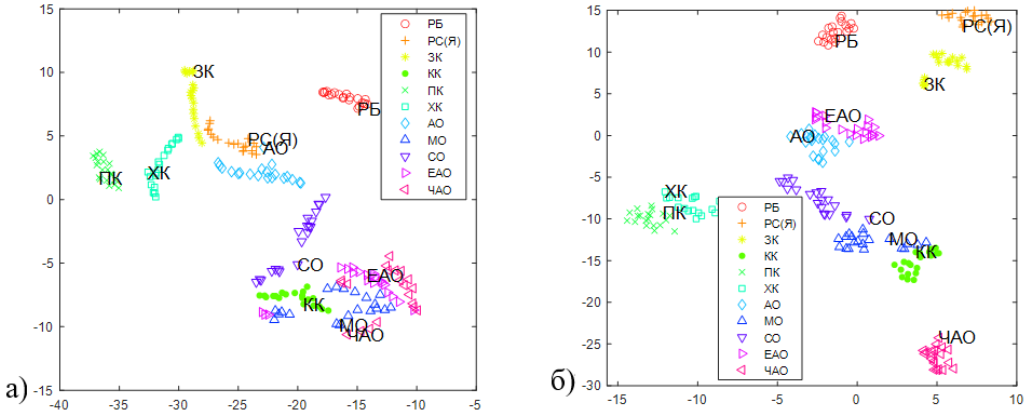

Рис. 7. Результат использования метода tSNE для реальных значений показателей (а) и ранговых значений показателей (б) демографии регионов ДФО.

Результат использования метода tSNE наглядно подтверждает возможность разделения регионов на классы. Географически соседние регионы оказались и соседями по результатам метода tSNE.

Кластеризация данных

В данной работе для классификации регионов, используются два метода: метод К-средних (K-Means) и самоорганизующаяся сеть (Self-orginizing map, SOM). Поскольку оба метода подразумевают задание количества классов разработчиком, то было принято решение делить данные на три класса, которые затем можно интерпретировать.

Метод К-средних стремится минимизировать суммарное квадратичное

отклонение точек кластеров от центров этих кластеров. По аналогии с методом главных компонент центры кластеров называются также главными точками. В результате разделение осуществляется так, чтобы изменчивость переменных внутри кластеров была малой, между кластерами – большой.

На рис. 8а показан результат работы метода К-средних в координатах метода tSNE. Класс 2 пересекается с классом 3, поэтому МО и СО принадлежат сразу двум классам. Это объясняется тем, что метод К-средних не гарантирует достижение глобального минимума суммарного квадратичного отклонения, а только одного из локальных минимумов [8].

Самоорганизующиеся сети являются усовершенствованной модификацией слоя конкурирующих нейронов («слоя Кохонена») [2]. В этой сети нейроны распределяются некоторым пространственным образом. Используя ранговые значения с помощью самоорганизующихся сетей, определяют принадлежность каждого региона ДФО к одному из трех классов.

На рис. 8б показан результат работы самоорганизующейся сети в координатах метода tSNE. Здесь регионы принадлежат, как правило, к одному классу, за исключением ЗК, который стоит на границе между 1 и 2 классами. Показательно, что ЧАО принадлежит к 1 классу, где находится и другой северный регион – РС(Я). Оба этих региона – самые северные, с депрессивными параметрами по демографии.

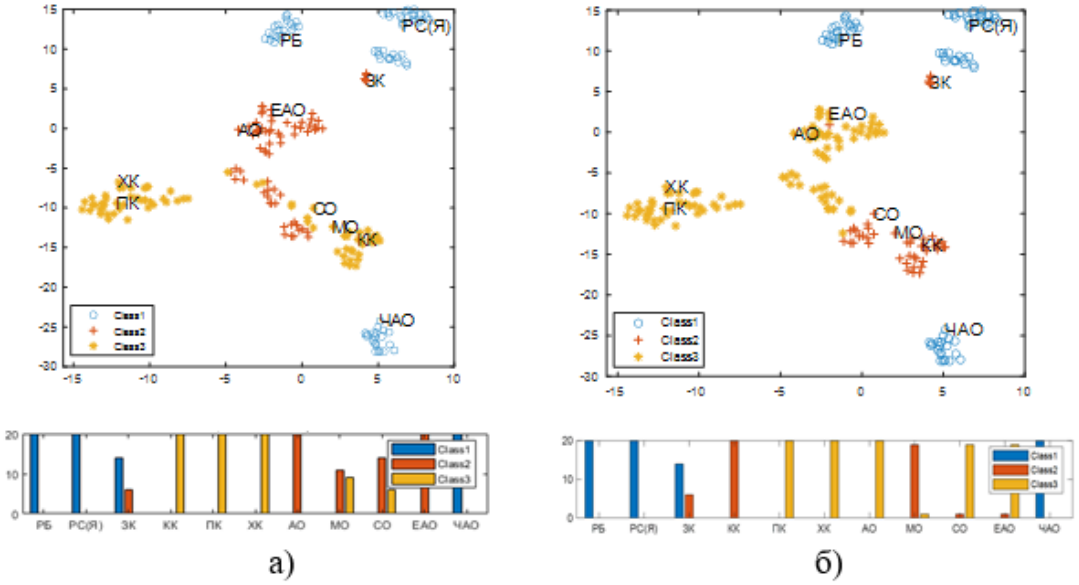

Рис. 8. Результат классификации регионов с помощью метода К-средних (а) и самоорганизующейся сети (б) по ранговым значениям демографических показателей регионов ДФО.

На основании географической карты регионов ДФО (рис.1) и результатов кластеризации (рис. 8) можно построить топографическую карту разделения регионов по демографическим показателям с помощью метода К-средних (рис. 9а) и самоорганизующейся сети (рис. 9б).

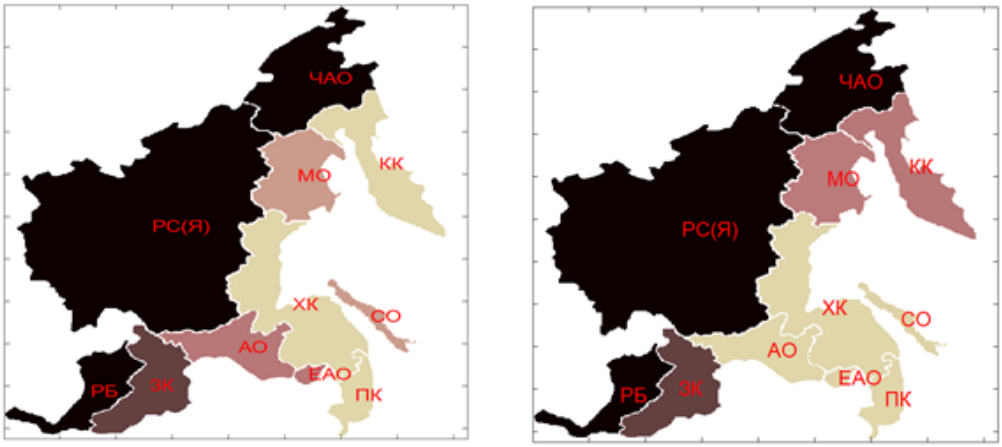

Рис. 9. Топографическая карта разделения регионов ДФО по демографическим показателям с помощью метода К-средних (а) и самоорганизующейся сети (б).

Существенным недостатком при такой классификации является необходимость заранее знать значения всех показателей по всем регионам ДФО в расчетный момент времени, чтобы пересчитать показатели в ранговые значения. Однако этим ограничением можно пренебречь, поскольку эти демографические показатели являются обязательными для расчета и публикуются Росстатом.

Оба подхода кластеризации получили соизмеримые решения, более 70% данных классияфицируются одинаково. Подход с самоорганизующимися сетями предпочтительнее, так как регионы принадлежащие классу 3, являются соседними «южными» регионами (АО, ЕАО, ХК, ПК, СО), с одинаковой демографической ситуацией. Наиболее депрессивные регионы (ЧАО, РС(Я), РБ, ЗК) попали в класс 1, а остальные регионы (ЗК, МО, КК) – в класс 2.

Поэтому класс 1 можно отнести к регионам с депрессивными демографическими показателями, класс 2 – к регионам с удовлетворительными демографическими показателями, а класс 3 – к регионам с хорошими демографическими показателями.

Заключение

В работе представлены демографические показатели регионов ДФО за 2000-2019 гг. Предложены методы предварительного анализа (столбчатые диаграммы и тепловые карты коэффициентов корреляции), которые позволяют оценить закономерности данных с последующим выбором методов машинного обучения.

Предложен метод замены показателей на ранговые значения и рассмотрено влияние такой замены на результат классификации.

Предложено наложение результатов классификации (методом К-средних и самоорганизующейся сетью) на данные tSNE, что позволило визуально оценить работу классификаторов. Применение двух различных подходов дает возможность верифицировать решение, так как в обоих случаях более 70% данных имеют одинаковый класс. По результату работы для классификации демографических показателей регионов ДФО выбрана самоорганизующаяся сеть, позволяющая определить, к какому классу относится регион: депрессивному, удовлетворительному или хорошему.

ЛИТЕРАТУРА

1. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining. – СПб.:Петербург, 2004.
2. *Безруков Н.С., Колосова Е.В.* Способы региональной кластеризации по параметрам человеческого капитала на основе самообучающихся нейронных сетей // Информатика и системы управления. – 2008. – №1(15). – С. 96-102.
3. *Сегаран Т.* Программируем коллективный разум. – СПб.: Символ-Плюс, 2008.
4. *Улумбекова Г.Э.* Здравоохранение России. Что надо делать. Состояние и предложения: 2019-2024 гг. – Изд. 3-е. – М.: ГЭОТАР- Медиа, 2019.
5. *Полянская Е.В., Безруков Н.С.* Система здравоохранения как один из факторов формирования человеческого капитала в Дальневосточном федеральном округе // Вестник ТОГУ. – 2020 – № 4 (59). – С.91-96.
6. *Калашников К.Н.* Ресурсное обеспечение российского здравоохранения: проблемы территориальной дифференциации // Экономические и социальные перемены: факты, тенденции, прогноз. – 2015. – № 1(37). – С. 72-85.
7. *Трибунский С.И. и др.* Типологизация субъектов Сибирского федерального округа на основе комплексной оценки здоровья населения, здравоохранения и социально–экономического развития // Сибирский медицинский журнал. – 2011. – Т. 26, № 4, вып. 1– С.175-178
8. *Mirkes E.M*. K-means and K-medoids applet. – University of Leicester, 2011.

Статья представлена к публикации членом редколлегии Ю.М. Перельманом.

E-mail:
Безруков Николай Сергеевич – bezrukow@mail.ru;
Полянская Елена Викторовна – dncfpd@dncfpd.ru.