УДК 314, 51-77, 519.6, 004-89
DOI: 10.25559/SITITO.14.201804.947-959

О ПОДХОДАХ К АНАЛИЗУ ДЕМОГРАФИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

А.И. Соловьев, С.А. Соловьев
Финансовый университет при Правительстве Российской Федерации, г. Москва, Россия

ON APPROACHES TO ANALYZING DEMOGRAPHIC DATA USING MACHINE LEARNING

Anatolii I. Solovev, Stefan A. Solovev
Financial University under the Government of the Russian Federation, Moscow, Russia

© Соловьев А.И., Соловьев С.А., 2018

Ключевые слова

Демография; распределение населения по возрастным группам; анализ данных; искусственный интеллект, машинное обучение, прогнозирование, экономическая эффективность.

Аннотация

Демографические данные являются достаточно доступными наборами данных, которые могут быть использованы для анализа с применением современных технологий искусственного интеллекта и машинного обучения (ML). Однако, они не могут быть использованы для этих целей без проведения специальных подготовительных процедур. К подготовительным мероприятиям относятся процедуры включающие работу с признаками, работу с отсутствующими данными, их нормализация и проектирования признаков. В статье на примере «Распределение населения по возрастным группам» показаны особенности демографических данных и предложены подходы по их подготовке к последующему использованию технологий искусственного интеллекта и машинного обучения для их анализа.

Проведенное исследования позволило получить следующие результаты. Установлено, что демографические данные обладают рядом особенностей, которые могут быть и должны использоваться в процессе повышения качества наборов данных для последующей работы с ними с использованием технологий искусственного интеллекта и машинного обучения. К особенностям демографических данных относится, в первую очередь, их временная упорядоченность, во вторых, демографические данные имеют предсказуемые пределы изменений, которые определяются социально-экономическими факторами, и отсутствие значительных перепадов между ближайшими значениями наблюдаемых данных.

На демографические данные оказывают влияние процессы в обществе социально-политического и экономического характера в различные исторические периоды, что необходимо учитывать при работе с демографическими данными. Демографическим данным, которые могут быть отнесены к определенным историческим периодам, следует уделять особое внимание поскольку их значения могут как повышать качество набора данных для машинной обработки, так и быть причиной возникновения и роста систематических и случайных ошибок.

Предлагаемые подходы могут иметь практическое приложение к решению задач прогнозирования народонаселения, определения структуры и состава возрастных групп, оценки продолжительности жизни, определение состава населения трудоспособного (экономически активного) возраста и ряда других задач.

Об авторах:

Соловьев Анатолий Ильич, кандидат технических наук, доцент, Департамент анализа данных, принятия решений и финансовых технологий, Финансовый университет при Правительстве Российской Федерации (125993, Россия, ГСП-3, г. Москва, Ленинградский пр., д. 49), ORCID: http://orcid.org/0000-0002-8544-7973, aisol@rambler.ru

Соловьев Стефан Анатольевич, аспирант, Финансовый университет при Правительстве Российской Федерации (125993, Россия, ГСП-3, г. Москва, Ленинградский пр., д. 49), ORCID: http://orcid.org/0000-0001-9443-1423, StefanSolovyev@me.com

Keywords

Demography; distribution of the population by age groups; data analysis; artificial intelligence, machine learning, forecasting; economic efficiency.

Abstract

Demographic data are fairly accessible data sets that can be used for analysis with the use of modern technologies of artificial intelligence and machine learning (ML). However, they cannot be used for these purposes without special preparatory procedures. Preparatory measures include procedures involving work with signs, work with missing data, their normalization and design of signs. The article on the example of "Distribution of the population by age groups" shows the features of demographic data and suggests approaches for their preparation for the subsequent use of artificial intelligence technologies and machine learning for their analysis.

The study allowed us to obtain the following results. It has been established that demographic data has a number of features that can be and should be used in the process of improving the quality of data sets for subsequent work with them using artificial intelligence and machine learning technologies. The features of demographic data include, first of all, their temporal ordering, secondly, demographic data have predictable limits of change, which are determined by socio-economic factors, and the absence of significant differences between the closest values of the observed data.

Demographic data is influenced by processes in a sociopolitical and economic society in different historical periods, which must be taken into account when working with demographic data. Demographic data that can be attributed to certain historical periods should be given special attention since their values can both improve the quality of the data set for machine processing and cause the occurrence and growth of systematic and random errors.

The proposed approaches can have a practical application to solving problems of population forecasting, determining the structure and composition of age groups, estimating life expectancy, determining the composition of the working (economically active) age population and a number of other tasks.

Введение

С настоящее время деятельность в сфере анализа данных приобретает особое значение. Анализ данных позволяет, обрабатывая большие массивы данных, выявлять в них статистические устойчивые зависимости, тенденции и взаимосвязи, которые могут быть использованы для повышения эффективности деятельности экономических агентов и экономического развития.

В нашу жизнь вошли такие понятия как большие данные (big data), открытые данные (public data). Ими в большом объеме владеют государственные органы и частные организации. Все больше данных появляется в открытом доступе, что привлекает широкие массы аналитиков (data analytics) для изучения данных, их анализа и использования результатов анализа для совершенствования различных сторон жизни общества и развития экономики. Все это стало возможным благодаря «взрывному» росту, распространению и использованию информационно-коммуникационных и компьютерных технологий, включая технологии искусственного интеллекта и машинного обучения. Благодаря им появились современные технологии, позволяющие обрабатывать огромные объемы данных в масштабе реального времени.

Исследования в сфере анализа данных с использованием современных технологий сегодня чрезвычайно востребованы. Но для эффективного использования наборов данных и получения практически значимых результатов исследователю необходимо глубоко понимать особенности данных с которыми ему предстоит иметь дело. От этого зависит целеполагание, постановка задач исследования и анализа данных, выбор методов и подходов к их обработке с использованием теоретико-вероятностных и статистических методов. Кроме этого, для применения современных методов анализа, основанных на технологиях искусственного интеллекта и машинного обучения, большое значение приобретает подготовка данных и обеспечение их надлежащего качества. Использование данных низкого качества является причиной получения недостоверных результатов и, как следствие, ошибочных решений. Это нашло отражения в афоризме «мусор на входе – мусор на выходе».

Вопросы исследования демографических процессов привлекают пристальное внимание в равной степени и политиков, и бизнесменов и ученых. В России внимание к вопросам демографии привлечено в связи с объявленной подготовкой к очередной пенсионной реформе, направленной на решение проблем пенсионной системы за счет повышения пенсионного возраста. Возникшая в обществе дискуссия демонстрирует широкий спектр взглядов. Но они являются по большей части субъективными мнениями, а не результатами исследований, имеющих научную базу с использованием современных технологий. Это побудило авторов обратить внимание на особенности демографических данных и предложить подходы для их анализа с использованием технологий искусственного интеллекта и машинного обучения. Алгоритмы обработки данных с использованием современных технологий построены таким образом, что позволяют обнаруживать в них закономерности, в том числе скрытые (неявные), т.е. те которые не могут быть обнаружены исследователем без использования машинных методов анализа данных.

Демографические данные Федеральной службы государственной статистики (ФСГС)[1] или Росстата России представля-

1 Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru (дата обращения: 23.09.2018).

ют собой обширный набор статистических данных о населении России, включая такие разделы как демография, уровень жизни, образование, здравоохранение, правонарушения, жилищные условия, неравенство и бедность, семья, материнство и детство, старшее поколение и положение инвалидов[2]. Эти статистические базы собраны, как правило, на основе периодических опросов в соответствии с существующими методическими документами[3] и их использование для анализа с использованием современных технологий не представляется возможным. В связи с этим, представляется актуальным проведение исследований и разработки подходов для анализа демографических данных с использованием современных технологий, основанных на технологиях искусственного интеллекта и машинного обучения.

Цель исследования

Настоящее исследование направлено на выявление особенностей демографических данных, их анализу и разработки подходов, позволяющих применить для анализа демографических данных современные методы анализа данных на основе технологий искусственного интеллекта и машинного обучения.

Исходными демографическими данными для проведения исследования взяты данные «Распределении населения по возрастным группам», представленные в разделе Демография на сайте ФСГС[4] (Таблица 1 на отдельном листе). В таблице представлены сведения о численности населения России (РСФСР[5] и Российская Федерация) в целом и по возрастным группам. Она представляет собой первичный набор данных. Для описания набора данных используют объекты и их признаки [1]. Объектами, в данном случае, выступают возрастные группы населения с интервалом 5 лет, а признаками – годы, в которые проводились наблюдения. Например, в первую группу включены дети от рождения до 4 лет включительно, а последняя группа содержит сведения о численности населения старше 70 лет. Наблюдения за численностью населения (переписи) осуществлялись с различной периодичностью. Набор данных содержит сведения о переписи населения начиная с 1926 года и далее 1935, затем 1959, 1970, 1979, 1989, и, наконец с 2002 по 2018 год включительно сведения о возрастном составе населения приведены ежегодно.

Одним из существенных недостатков, препятствующих анализу данных с использованием современных технологий искусственного интеллекта и машинного обучения, является наличие пропусков данных. Этот недостаток в полной мере присутствует в анализируемом наборе данных. В первую очередь, это связано с тем, что наблюдения за численностью населения осуществлялись периодически и сведения по годам, в которые не проводились наблюдения в наборе данных отсутствуют. Кроме этого, отсутствие данных связано с рядом исторических периодов советской и российской истории, такими как Великая Отечественная война 1941 – 1945 годов и реформы 1990 – 2001 годы, в которые происходила серьезная социально-политическая трансформация в жизни общества. С 2002 года совершенствование деятельности статистических служб позволило вести ежегодные наблюдения за составом населения, что позволяет определить набор данных за этот период как наиболее подходящий для применения машинных методов анализа данных. Однако, в этом случае требуется оценка достаточности объема этой выборки для её применения при решении задач кратко- и долгосрочного прогнозирования.

Для достижения поставленной цели исследования необходимо устранения имеющихся в наборе данных недостатков и подготовки их для машинного анализа. Это может быть достигнуто в результате решения следующих задач:

1. Анализ демографических данных и определение свойственных им особенностей.

2. Разработка подходов и методов повышения качества демографических данных с целью их последующей машинной обработки и анализа с использованием современных технологий.

Таблица 1. Распределение населения России по возрастным группам (тыс. чел.)
Table 1. Distribution of the population of Russia by age groups (thousand people)

Возрастные группы\годы	1926	1939	1959	1970	1979	1989	2002	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Все население	**92681**	**108377**	**117534**	**129941**	**137410**	**147022**	**145167**	**144134**	**143801**	**143236**	**142863**	**142748**	**142737**	**142857**	**142865**	**143056**	**143347**	**143667**	**146267**	**146545**	**146804**	**146880**
0-4	14114	13806	13353	9326	10523	12032	6399	6660	6916	7066	7234	7433	7671	7968	8051	8380	8687	8899	9262	9512	9582	9347
5-9	9420	11735	12415	11975	9707	11360	6941	6762	6583	6511	6503	6638	6783	7091	7117	7261	7441	7662	8004	8218	8558	8873
10-14	10994	14158	8501	13202	9512	10592	10406	9314	8604	7940	7458	7056	6891	6610	6601	6567	6689	6823	7126	7254	7408	7598
15-19	10947	9495	8975	12291	12385	9968	12801	12544	12212	11852	11244	10485	9650	8389	8237	7631	7152	6956	6829	6731	6690	6816
20-24	8732	8744	11552	9706	12995	9755	11466	11870	12081	12098	12298	12457	12389	12169	12122	11599	10849	9971	9293	8445	7828	7336
25-29	7324	10454	10591	7102	11902	12557	10613	10797	10879	11054	11130	11358	11667	11982	12012	12328	12556	12522	12620	12412	11879	11120
30-34	5420	8820	11103	11708	8016	12863	9836	10030	10228	10316	10466	10537	10696	10980	11016	11116	11346	11660	12092	12219	12537	12766
35-39	5171	7240	6423	9327	8399	11684	10216	9665	9416	9427	9485	9705	9885	10172	10211	10380	10459	10614	10884	11098	11194	11425
40-44	4348	5315	6177	10925	10485	7663	12546	12155	11641	10925	10325	9800	9409	9241	9251	9340	9563	9750	10122	10220	10381	10453
45-49	3790	4268	7167	6698	9376	7955	11606	11891	11906	12070	12084	11929	11634	10672	10561	10023	9545	9187	9140	9193	9280	9499
50-54	3219	3710	5965	5253	9716	9593	10071	10447	10576	10738	10887	11037	11272	11483	11509	11560	11436	11184	10957	10356	9835	9372
55-59	2787	3332	4751	6874	5596	8399	5347	6466	7737	8724	9164	9501	9755	10022	10063	10215	10382	10634	10873	11093	11155	11049
60-64	2430	2775	3590	5510	5065	8360	7983	6387	5213	4458	4408	5014	5916	7832	7982	8380	8690	8949	9260	9445	9610	9783
65-69	1721	2079	2664	4181	5492	4510	6345	7021	7567	7699	7572	6687	5565	4002	3913	3896	4453	5269	6428	7263	7637	7937
70 и более	2212	2426	4303	5806	8200	9646	12469	12325	12242	12358	12605	13111	13554	14210	14219	14380	14099	13587	13377	13086	13230	13506

Источник: ФСГС, 2018.

2 Население // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/ (дата обращения: 23.09.2018).

3 Методологические пояснения // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/free_doc/new_site/population/demo/metod/met-dem.htm (дата обращения: 23.09.2018).

4 Демография // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/# (дата обращения: 23.09.2018).

5 Российская Советская Федеративная Социалистическая Республика – государственное название России в период с 1922 по 1991 годы.

Основная часть

Особенности демографических данных и их характеристики

Любые наборы данных обладают рядом особенностей, знание которых позволяет устранять присущие этим наборам данных недостатки и повышать качество данных, т.е. подготовить их для машинной обработки и анализу с использованием современных технологий.

В книге «Машинное обучение» [2] так определены 5 преимущества применения машинного обучения:

1. *Точность*. Машинное обучение использует данные для создания принимающей решение программы, оптимизированной под поставленную задачу. По мере накопления данных автоматически возрастает точность прогнозов.

2. *Автоматизация*. По мере подтверждения и отбрасывания ответов MLмодель (модель машинного обучения) может автоматически обнаруживать новые шаблоны. Это позволяет встраивать машинное обучение непосредственно в автоматизированные рабочие процессы.

3. *Скорость*. Машинное обучение дает ответы за доли секунды после поступления новой информации, позволяя системам реагировать в реальном времени.

4. *Возможность настройки*. Многие задачи, управляемые данными, можно решить с помощью машинного обучения. Модели строятся на базе ваших собственных данных и допускают настройку под любую систему мер, принятую в вашем бизнесе.

5. *Масштабируемость*. При росте бизнеса MLмодель легко приспосабливается к увеличивающимся объемам данных. Некоторые алгоритмы можно использовать для обработки множества данных на разных вычислительных машинах в облаке» [2, с. 42].

Но чтобы реализовать указанные выше преимущества для машинного обучения должны использоваться качественные наборы данных.

Для анализа данных с использованием машинного обучения определяющее значение имеет качество обучающей выборки, которая используется для обучения ML-модели (модели машинного обучения), построенной на базе искусственного интеллекта. Качество исходных данных, используемых для формирования обучающей выборки, является фундаментом машинного обучения.

Основными этапами подготовки данных для машинного обучения являются:

- сбор данных;
- определение входных признаков;
- определение целевой переменной;
- определение достаточности объема и репрезентативности обучающей выборки;
- подготовка данных к моделированию.

Сбор данных. На этапе сбора данных важно определить не только базовый набор данных, который планируется использовать в качестве обучающей выборки для построения ML-модели, но и предусмотреть сбор вспомогательных сведений, которые могут быть полезны для повышения качества обучающей выборки на этапе подготовки данных.

В настоящем исследовании основным набором данных являются сведения, приведенные в таблице 1 «Распределении населения по возрастным группам». Объектами являются возрастные группы населения с интервалом 5 лет, а признаками – годы, в которые проводились наблюдения.

Для повышения качества этого набора данных могут потребоваться дополнительные сведения из источников как входящие в состав статистических баз данных (ЦБСД[6], ЕМИСС[7]) и публикациях Росстата России, например Демографический сборник России[8], так и сведения, которые могут быть найдены на информационных ресурсах в сети Интернет, в том числе демографические данные международных статистических баз таких как Специальный стандарт распространения данных Международного валютного фонда (ССРД МВФ)[9] и Открытые данные Всемирного банка (WBOD)[10]. Вспомогательные сведения использованные в рамках настоящего исследования будут представлены в разделах где они были востребованы при решении поставленных задач.

Определение признаков. На этапе определения признаков необходимо ответить на вопрос о достаточности признаков с наборе данных. Ответ на этот вопрос может предполагать как исключение ряда признаков из набора данных так и проектирование данных, т.е. определение и дополнение набора данных признаками, которые позволяют повысить качество обучающей выборки. Как правило это определяется тем какое влияние могут оказывать признаки на достижение поставленной цели. Чем меньше степень влияния признака на целевой результат тем больше вероятность того, что его следует исключить из набора данных, т.к. этот признак будет снижать качество обучающей выборки. И наоборот, разработка дополнительных признаков имеющих высокую корреляцию с целевым параметром, будет способствовать повышению качества обучающей выборки. Нередко, решение об изменении состава признаков может приниматься на этапе оценки точности ML-модели, но это входит непосредственно в процесс машинного обучения, что в структуру настоящего исследования не входит.

Определение целевой переменной. Определение целевой переменной, как правило, имеет место при решении задач прогнозирования с использованием ML-модели. В настоящем исследовании целевой переменной является численность населения соответствующей возрастной группы. Это определяется тем, что прогнозы численности населения по возрастным группам позволяют путем агрегирования данных по различным социальным группам получить прогнозные сведения, например, по группе населения от 5 до 15 лет, что может быть использовано при оценке потребности мест в образовательных учреждениях начального и среднего образования. Аналогичные сведения могут быть агрегированы в прогнозы численности по группам

6 Федеральная служба государственной статистики [Электронный ресурс]. URL: http://cbsd.gks.ru (дата обращения: 23.09.2018).
7 Единая межведомственная информационно-статистическая система (ЕМИСС) [Электронный ресурс]. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/databases/emiss/ (дата обращения: 23.09.2018).
8 Демографический ежегодник России - 2015 г. // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/bgd/regl/B15_16/Main.htm (дата обращения: 23.09.2018).
9 Special Data Dissemination Standard. Russian Federation Population // International Monetary Fund [Электронный ресурс]. URL: https://dsbb.imf.org/sdds/dqafbase/country/RUS/category/POP00 [Электронный ресурс]. URL:
10 World Bank Open Data [Электронный ресурс]. URL: https://data.worldbank.org (дата обращения: 23.09.2018).

населения трудоспособного возраста или населения, которое будет нуждаться в пенсионном обеспечении для планирования фондов денежных средств Пенсионного фонда России (ПФР).

Определения достаточности объема и репрезентативности обучающей выборки так же является важным подготовительным этапом.

Определение достаточности объема выборки. Ответ на вопрос о достаточности объема обучающей выборки во многом зависит от конкретной задачи, которую предстоит решить. Универсальных рекомендаций на этот случай нет. Поэтому рассмотрим основные факторы, которые следует учитывать на этом этапе:

- сложность решаемой задачи. Чем сложнее решаемая задача, тем больше данных будет необходимо для того, чтобы с высокой степенью достоверности могла быть выявлена связь между признаками и целевой переменной. И наоборот, если основные связи носят линейный характер, то объем выборки может быть уменьшен;

- требования к точности. Если требования к точности невелики и достаточно, например, 60% успешных результатов прогнозирования, то можно ограничиться обучающей выборкой меньшего объема, чем в случае, если необходимо получить до 95% успешных результатов;

- размерность (количество) входных признаков в наборе данных. Если количество признаков невелико, то объем обучающей выборки может быть меньше. И наоборот, при большом числе входных признаков потребуется обучающая выборка большого объема, позволяющая в достаточной степени выявить закономерности и связи между признаками.

Набор данных, принятый в качестве исходного набора, обладает, в этом плане, следующими характеристиками:

- ограниченность данных. Данные представленные в таблице 1 не могут быть отнесены к категории «больших данных», что не позволяет осуществлять процедуры повышения (снижения) размерности или увеличения (снижения) объема данных без использования методов их дополнительной обработки, например, проектирование признаков, которые будут использованы далее. Поэтому в настоящем исследовании набор данных используется в полном объеме и решения о возможности изменения объема обучающей выборки будут приниматься с учетом результатов, получаемых в ходе исследования на основе выявленных особенностей демографических данных и возможностей их использования для повышения качества исходного набора данных, включая и решения по изменению объема обучающей выборки;

- малая размерность. В наборе данных представлена только численность населения по возрастным группам по году наблюдения. Поэтому для повышения качества данных необходимо исследовать вопросы связанные с использованием проектирования дополнительных признаков;

- отсутствие показателей и их параметров, позволяющие оценивать точность работы модели. Они так же должны быть определены по результатам исследования. В качестве предварительной оценки можно определить, что при работе с демографическими данными требуется достаточно высокая точность получаемых результатов, которую в виду отсутствия внешних критериев целесообразно установить на уровне не менее 85% успешных результатов прогнозирования. Исходя из этого на первом этапе обучения модели может быть определена достаточно или нет того объема выборки, которые представлены в исходном наборе данных.

Определения репрезентативности выборки. Кроме размера обучающей выборки важным фактором, влияющим на точность прогнозов MLмодели, является *репрезентативность* выборки. Оценка репрезентативности ориентирована на то, чтобы оценить насколько входящие в набор данных признаки похожи на те, которые будут накапливаться в будущем. Крайне важно, чтобы обучающая выборка включала в себя признаки тех типов, для которых будет создаваться прогноз. Выборка, содержащая признаки, не имеющие отношение к будущим данным, будет способствовать формированию и росту *систематической ошибки отбора* (ковариационный сдвиг).

Отсутствие репрезентативности в обучающей выборке, как правило, связано со следующими факторами:

- если получение экспериментальных значений целевой переменной возможно только для определенного, содержащего ошибку подмножества данных;

- если свойства признаков претерпевали со временем изменения;

- если набор признаков изменяется со временем.

Чтобы избежать подобных проблем, обучающая выборка должна максимально репрезентативно представлять будущие данные.

Исходный набор данных (Табл. 1) настоящего исследования содержит малое количество признаков. Важно отметить, что по своему характеру они идентичны тем признакам целевой функции, которые будут предметом прогнозирования с использованием машинного обучения. При этом набор признаков также не будет изменяться. Это вполне удовлетворяет первому признаку репрезентативной модели.

Данные признаки сформированы по методикам, которые использует Росстат России в течении продолжительного времени и есть основания полагать, что с течением времени их свойства не претерпят значительных изменений. Подтверждением этого может являться тот факт, что Росстат России на своем сайте по существующим методикам[11] осуществляет прогнозирование демографических факторов, а именно, состава и структуры населения Росси до 2035 года включительно[12]. Однако, он включает только 3 социальные группы граждан: моложе трудоспособного возраста, трудоспособного возраста и старше трудоспособного возраста. Это вполне удовлетворяет второму признаку репрезентативной модели.

Подготовка данных. Подготовка данных к моделированию представляет собой основной этап нашего исследования и предполагает повышение качества исходного набора данных на основе выявления и последующего использования характерных особенностей демографических данных.

Процедура подготовки данных, как правило, включает в себя ряд процедур таких как работа с признаками, работа с отсутствующими данными, нормализация, чтобы сделать призна-

11 Методологические пояснения // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/free_doc/new_site/population/demo/metod/met-dem.htm (дата обращения: 23.09.2018).

12 Демографический прогноз до 2035 года (по состоянию на 22.02 2018 г.) / Демография // Федеральная служба государственной статистики [Электронный ресурс]. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/# (дата обращения: 23.09.2018).

ки пригодными для сравнения и исключить ошибки разномасштабности признаков, и проектирования признаков.

В силу особенностей исходного набора данных исследования по подготовки данных были сгруппированы следующим образом. Первая группа включила работу с признаками и их нормализация, а вторая группа – работу с отсутствующими данными и проектирование признаков. Это связано с тем6 что именно вторая группа процедур представляет собой основной объект настоящего исследования и его результаты будут рассмотрены с следующем разделе статьи.

Работа с признаками и их нормализация. Как правило при работе с данными приходится иметь дело с числовыми и категориальными признаками. При этом наибольшую сложность представляет собой работа с категориальными признаками. Однако в нашем случае в исходном наборе данных нет признаков, имеющих категориальный характер. Поэтому работа с признаками ограничивается констатацией того факта, что все признаки являются численными и не требуют проведения с ними дополнительных действий.

Нормализация признаков означает, что каждый признак обрабатывается с целью его приведения к единой шкале. Диапазон значений признака может влиять на его важность относительно других признаков. Если значения одного признака варьируются от 0 до 10, а второго — от 0 до 1, вес первого признака по отношению ко второму составит 10 и соответственно увеличит его влияние на целевую функцию. Иногда целесообразно вес конкретного признака регулировать вручную, а можно оставить определение относительных весов признаков на откуп MLалгоритму. Но чтобы гарантировать одинаковый подход к оценке влияния признаков то данные требуется нормализовать. Зачастую нормализация приводит данные к единому диапазону значений, например, к диапазону значений от 0 до 1 или от –1 до 1.

Исходя из особенностей исходного набора данных можно отметить, что все признаки соответствуют единой шкале измерения – тыс. человек, что не требует применение к нему процедуры нормализации.

Но в отличии от работы с признаками процедуры нормализации, работы с отсутствующими данными и проектирование данных могут быть связаны с необходимостью сбора информации из дополнительных источников и/ или преобразования данных исходного набора и данных из дополнительных источников с целью восстановления отсутствующих значений и проектирования новых признаков, использование которых для целей прогнозирования может потребовать проведение процедуры нормализации. Поэтому к процедуре нормализации следует при необходимости вернуться после проведения процедур по работе с пропущенными данными и проектирования признаков.

Основные исследования по выявлению особенностей демографических данных и использования этих особенностей имеют место при работе с отсутствующими данными и проектировании признаков. Результаты этих исследований представлены в следующем разделе.

Полученные результаты

Одним из основных недостатков исходного набора данных являются значительные пропуски и отсутствие данных. Важную роль при работе с отсутствующими демографическими данными играет то, что демографические данные являются яркими представителями временных рядов. Их основное свойство заключается в том, что в них присутствует временная упорядоченность, что можно и нужно использовать при работе с отсутствующими данными.

Особенности демографических данных. Как правило, отсутствие данных в наборах носит частичный характер, т.е. отсутствуют отдельные значения признаков, которые измерить не представилось возможным и они с определенной долей вероятности могут быть восстановлены с учетом рядом расположенных данных. Нередко само отсутствие данных несет в себе скрытую информацию, которая может быть использована ML-алгоритмом. Такие отсутствующие данные называются информативными.

Как было отмечено выше демографические данные являются временными рядами, т.е. в них присутствует временная упорядоченность.

Следующей особенностью демографических данных является то, что в силу практики ведения статистического учета, измерение демографических признаков осуществлялось периодически и т.о. значительные временные периоды не имеют измеренных значений признаков и представляют собой отсутствующие данные. В нашем исходном наборе такие периоды, не имеющие измеренных значений признаков объектов включают в себя период до 1926 года, период 1927-1938 год, 1940-1958 год, 1960-1969 год, с 1971 по 1978 год, 1980 по 1989 годы и с 1990 до 2001 года. С 2002 года статистические наблюдения велись в ежегодном режиме, что не требует применение к данному периоду работ с пропущенными данными.

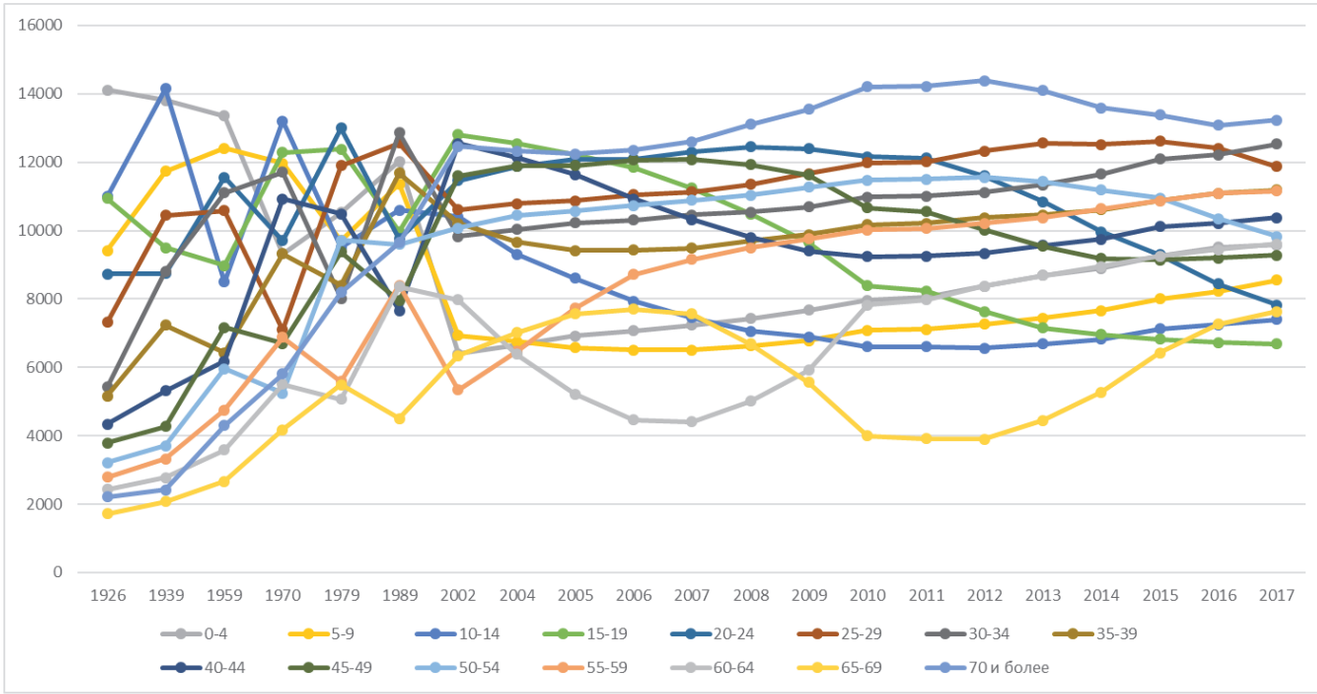
Одним из подходов позволяющих провести работу с отсутствующими данными является визуализация. Поэтому был построен график исходных данных, который представлен на рисунке 1.

При построении графика были использованы данные по 2017 год включительно. Это вызвано тем, что на момент проведения исследования данные по 2018 году могут рассматриваться как текущие значения и не являются окончательными. Их следует включить в рассмотрение после завершения 2018 года и подведения окончательных итогов года.

Построенный график позволяет по иному взглянуть на тенденцию и динамику изменения численности населения различных возрастных групп за рассматриваемый период. Очевидно, что график, в представленном виде, имеет существенный недостаток, а именно разномасштабность т. к. отсутствующие данные на нем не отображаются и приводят к неверному представлению графических данных.

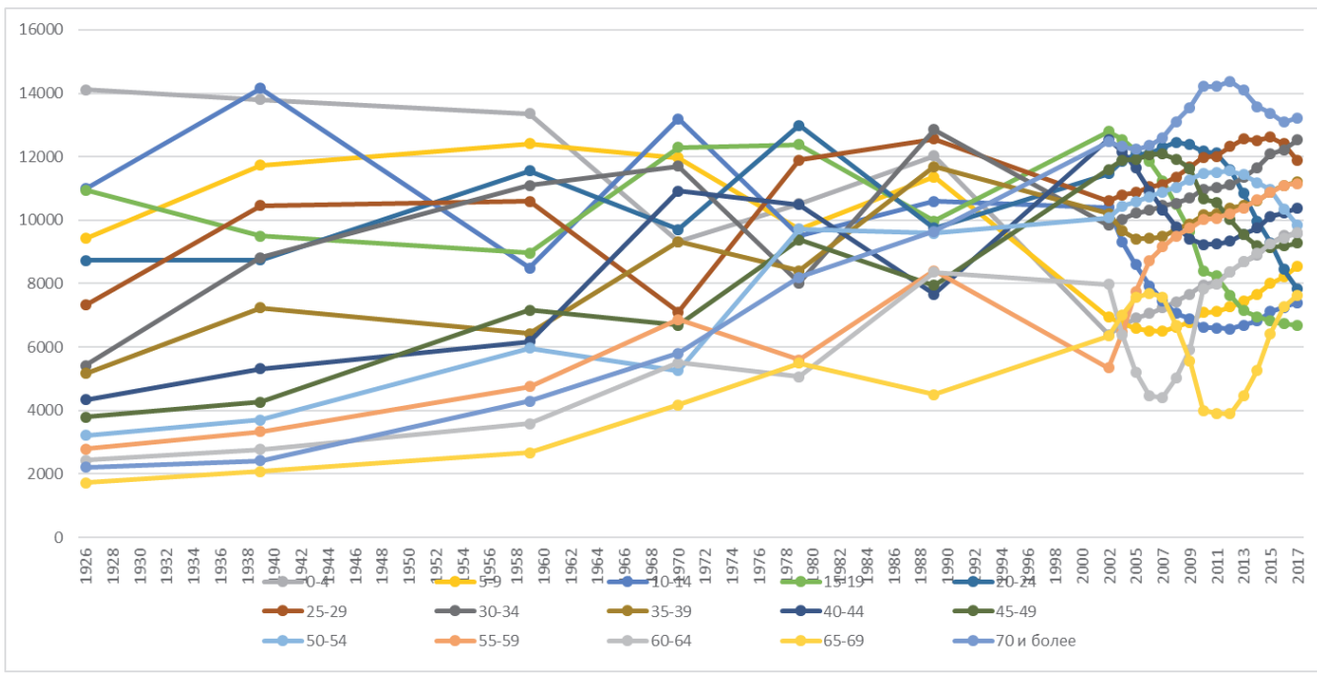
Для устранения этого недостатка были в набор исходных данных добавлены недостающие столбцы с внесенными в них годовыми метками, а используя возможности Microsoft Exel был использована опция при которой отсутствующие данные игнорируются и соединяются в график только значимые показатели.

Визуализация исходного набора данных позволяет выявить первую особенность демографических данных – **изменение данных носит последовательный характер, последующие значения имеют сильную связь с предыдущими и не могут иметь разрывов или скачкообразных изменений**. Все изменения в составе возрастных групп, как правило, определяются такими демографическими факторами как рождение, смерь и миграция, которая может иметь существенное значение.



Источник: авторы, 2018.

Рис. 1. График изменения численности населения по возрастным группам с 1926 по 2017 г.г.
Fig. 1. Graph of population change by age groups from 1926 to 2017



Источник: авторы, 2018.

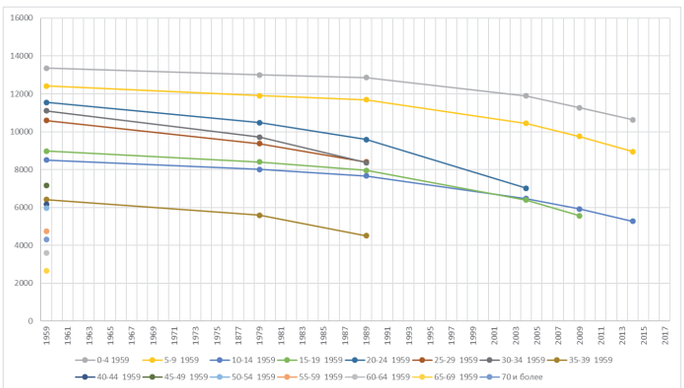
Рис. 2. График изменения численности населения по возрастным группам с 1926 по 2018 г.г. (равномерная шкала)
Fig. 2. Graph of population change by age groups from 1926 to 2018 (uniform scale)

Например, миграция могла увеличиваться в годы Великой Отечественной войны в связи с большим количеством перемещенных лиц в эвакуацию в другие республики СССР и лицами, которые были угнаны на работы в Германию. Данные о количестве перемещенных лиц могут быть частично восстановлены по историческим источникам, но в рамках нашего исследования это не требуется поскольку восстановить эти данные в разрезе возрастных групп не представляется возможным. Другой период серьезного влияния миграции на состав возрастных групп имел место в период с 1991 по 2000 годы, когда после развала СССР началось перемещение русскоговорящего населения из бывших республик СССР на территорию современной России, а также увеличение иммиграции за пределы России в Ближнее и Дальнее зарубежье. Однако, как показали дальнейшие исследования относительные изменения состава возрастных групп от миграции носят незначительный характер и в дальнейшем особо анализироваться не будут. Кроме этого, современные статистические данные по вопросам миграции не имеют характеристики по возрастным группам, что делает их малопригодными для целей настоящего исследования.

Особенно явно эта закономерность проявляется в период проведения регулярных наблюдений в период с 2002 по 2018 годы, где состав большинства возрастных групп изменяется плавно, что может быть использовано при проектировании признаков на последующем этапе исследования. Обращают на себя только изменения в составе трех возрастных групп: «60-64», «65-69» и «свыше 70». Первые две группы за данный период претерпевали значительные изменения в своем составе. По группе «60-64» за данный период отмечалось снижение с 7083 тыс. чел. в 2002 году до 4408 тыс. чел. в 2007 году, т.е. почти в 2 (1,81) раза за 5 лет с последующим восстановление численности к 2011 году. По группе «65-69» имело место снижение численности с 7699 тыс. чел в 2006 году до 3896 тыс. чел в 2012, т.е. также почти в 2 (1,97) раза за 6 лет. Следует заметить, что данные возрастные группы складывались из поколения людей родившихся в период 1942 – 1952 годы, которые в истории страны связаны с военными действиями и послевоенным восстановлением страны, что сопровождалось снижением рождаемости. Следовательно снижением численности рассмотренных в первую очередь определяется тем, что от рождения численность людей принадлежащих к данному поколению было меньше и результаты наблюдения связаны с таким явлением как «демографическая яма». Аналогичная «демографическая яма» имела место в период 90-х годов, когда рождаемость в стране упала и состав возрастной группы «0-4» снизился с 12032 тыс. чел до 6399 в 2002 году, т.е. в 1,88 раза. Следует отметить, что динамика изменения состав возрастных групп в период с 1989 по 2002 год практически отсутствуют и в ходе исследования будут предложены подходы по работе с отсутствующими данными за указанный период и их проектированию.

Как отмечено выше, обращает на себя внимание рост численности группы «70 и выше». Эта группа является особенной. Во-первых, это определяется тем что она включает в себя, в отличии от других групп, не 5-ти летний интервал наблюдения, а всех представителей населения достигших возраста 70 лет. Интересно, что поколение родившееся в 1926 году достигла возраста 70 лет в 1996 году, а в настоящее время еще живы представители этого поколения, возраст которых составляет 92 года[13]. Таким образом, в численность группы «70 и свыше» формируется поколениями за весь период наблюдения с 1926 года и не случайно, что эта возрастная группа является наиболее многочисленной.

Особенность демографических данных, выявленная выше, позволила предположить подход который заключается в построении «Линий поколения», т.е. исследовать изменения происходящие от момента рождения и далее на примере возрастных групп. Если в 2002 году возрастная группа «0-4» объединяла в своем составе лиц от рождения до 4 лет, то через 5 лет группа «5-9» будет включать в себя преимущественно тех же лиц, но повзрослевших на 5 лет и т.д.



Источник: авторы, 2018.

Рис. 3. Графики «Линий поколений» по данным возрастных групп изменения с 1959 по 2017 г.г. (равномерная шкала)

Fig. 3. "Generation Line" plots according to the age group of changes from 1959 to 2017 (uniform scale)

На рисунке 3 представлены «линии поколений», которые удалось построить по имеющимся данным с 1959 по 2017 г. Особенностью построения «линий поколений» состоит в том, что в имеющихся данных не все признаки могут быть использованы для их построения. Например, для их построения мы можем без серьезной ошибки использовать данные возрастных групп с периодом не менее5 лет и кратно 5 годам. Так для построения «линии поколения» граждан возрастной группы «0-4» по данным 1959 года, были использованы данные возрастной группы «20-24» по данным 1979 г., группы «30-34» по данным 1989 г. и далее группы «45-49» по данным 2004 г., групп «50-54» - 2009 г. и группы «55-59» по данным последнего в этой последовательности 2014 г. Данная логика соответствует графику «0-4 1959» на рисунке 3. Таким образом в обозначении графика возрастная группа указана по состоянию на 1959 год. Аналогичные графики могут быть построены и для других поколений, данные которых позволяют сопоставить 5-ти летний размер группы и кратный 5 интервалы наблюдений.

Анализ «Линий поколений показывает, что линии основным фактором влияющим на изменение состава соответствую-

13 «Сенатор В. Долгих является старейшим в Совете Федерации. Его возраст 93 года». Чуракова О., Мухаметшина Е. Как живут непростые российские пенсионеры // Ведомости. 2018. 28 августа. [Электронный ресурс]. URL: https://www.vedomosti.ru/politics/articles/2018/08/29/779257-zhivut-neprostie (дата обращения: 23.09.2018).

щей возрастной группы анализируемого поколения является смертность независимо от причин. В своем исследовании мы ввели понятие «Коэффициент выбытия», который показывает за какой период и насколько произошло снижение состава возрастной группы. Эти данные могут быть использованы для расчета отсутствующих данных по этой возрастной группе в периоды, когда данные наблюдения недоступны. Коэффициент выбытия также учитывает изменения в составе групп связанные не только в естественной убылью граждан, но и учитывает выбытие граждан выезжающих на постоянное место жительства за пределы Российской Федерации.

Коэффициент выбытия рассчитывается по формуле

Коэффициент выбытия = (Численность группы текущая (0) – Численность группы (- N лет)) (1)

Численность группы(- N лет)

где N – временной интервал кратный 5 годам.

Таблица 2. Значения коэффициента выбытия для различных возрастных группы за период 5, 10 и 15 лет (за период 2004 – 2018 годы)
Table 2. Values of the retirement rate for different age groups for the period of 5, 10 and 15 years (for the period 2004 - 2018)

Группы	Коэффициент выбытия		
0-4	За 5 лет	За 10 лет	За 15 лет
5-9	0,021		
10-14	0,021	0,022	
15-19	0,019	0,027	0,023
20-24	0,026	0,040	0,085
25-29	0,025	0,061	0,194
30-34	0,017	0,025	0,018
35-39	0,007	0,006	-0,037
40-44	-0,001	-0,008	-0,032
45-49	-0,007	-0,021	-0,053
50-54	-0,018	-0,044	-0,030
55-59	-0,034	-0,074	-0,091
60-64	-0,058	-0,114	-0,177
65-69	-0,087	-0,165	-0,240

Источник: авторы, 2018.

При уменьшении численности возрастной группы Коэффициент выбытия имеет отрицательное значение, поскольку в естественных условиях состав возрастной группы имеет тенденцию к уменьшении численности ввиду существующей смертности по естественным или неестественным причинам. Если по причинам связанным в первую очередь с процессами миграции и ростом состава возрастной группы его показатель может принимать положительное значение. В настоящее время в возрастной группе «50-54» за 5 лет коэффициент выбытия имеет значение менее -2%, а за 15 летний период -3%. То для возрастной группа «65-69» снижение численности достигает 25% на интервале 15 лет.

Положительные значения Коэффициента выбытия свидетельствует о том, что за указанный период имел место рост численности возрастной группы. Это, как правило, является следствием миграционных явлений, когда пополнение в составе групп происходи не за счет естественной рождаемости, а за счет приезжающих на постоянное место жительства в страну граждан других государств. Важно отметить, что если для младших возрастных групп в настоящее время миграционное движение позволяет компенсировать депопуляцию групп, то для старших возрастных групп миграционные явления не могут компенсировать их естественное выбытие.

Таким образом, визуализация данных позволяет выявить ряд особенностей демографических данных, которые имеют значение для подготовки их к машинной обработке и повышений их качества.

Для полноты картины следует обратить внимание, что демографические данные тесно связаны с процессами политической и социально-экономической жизни общества (государства). Поэтому при определении роли и значения отдельных группы демографических данных следует дополнительно сопоставить их с историческими, социально-экономическими и политическими процессами, которые происходили в стране в рассматриваемые исторические периоды.

С этой точки зрения можно выделить несколько исторических периодов, которые до сих пор оказывают влияние на формирование демографических данных в Российской Федерации:

- период 1941 – 1960 г.г. включающий в себя годы Великой Отечественной войны и послевоенного восстановления народного хозяйства;

- период 1991 – 2000 годов когда в силу развала Советского Союза и социалистического строя, т.е. смены государственного устройства, была нарушена общественная жизнь всех слоев общества, были разрушены экономические устои и главенствовали процессы перехода от социалистической к рыночной социально экономической системе.

Демография этих периодов содержит в себе явления, которые следует признать неестественными для жизни обычного общества и рассматривать их как эксцессы (выбросы) и принимать решение об использовании данных таких периодов с учетом целей и задач, которые поставлены исследователем в рамках проводимой работы.

В настоящем исследовании мы готовим данные для их последующей машинной обработки и прогнозирования состава населения по возрастным группам ориентировочно до 2035 года. Согласно Стратегии социально-экономического развития России в этот период намечено планомерное решение задач направленных на социально экономическое развитие страны и повышение благосостояния её граждан. Исходя из этого. наиболее репрезентативным периодом является период с 1960 по 1990 г.г. Данные период отличается поступательным развитием общества, последовательными изменениями демографического состава, что может являться существенной дополнительной информацией для проектирования данных.

Выявленные особенности демографических данных позволяют перейти непосредственно к работе с отсутствующими данными и их проектированию.

Подходы к работе с отсутствующими данными. Как было отмечено ранее отсутствующие данные могут иметь различный характер:

- информативный;

- не информативный.

Если отсутствующие данные имеют информативный характер, анализ пропущенных значений позволяет выдвинуть гипотезу о значении и величине отсутствующих данных и использовать её при проектировании, заменяя отсутствующие

данные проектируемыми с учетом их информативности. В нашем исследовании демографические данные имеют не информативный характер поскольку являются результатами наблюдений.

Если данные носят не информативный характер, то это не позволяет выдвинуть гипотезу об их предполагаемом значении и замена их без дополнительного анализа будем снижать качество набора данных и их использования для машинного обучения.

Для работы с неинформативными отсутствующими данными существует ряд подходов:

- исключение из набора данных строк или столбцов содержащих отсутствующие данные. Такой подход неплохо работает в случае когда имеет место большой набор данных и исключение некоторого количества строк или столбцов (объектов или признаков) не оказывает серьезного влияния на качество набора данных и делает его пригодным для машинной обработки.

В нашем случае набор данных недостаточно велик и исключение из него значительного объема данных, например, за период с 1990 по 2001 годы, приведет к тому, что набор демографических данных будет недостаточного объема или нерепрезентативным. Кроме того, отсутствующие данные не носят случайный характер, что в случае их простого исключения из набора данных может вести к росту систематической ошибки при последующей машинной обработке.

- использование временной упорядоченности. Как было установлено, демографические данные обладают свойством временных рядов и это свойство должно быть в полной мере использовано для реализации данного подхода.

- наличие статистических закономерностей. В этом случае, отсутствующие данные могут быть без опасения внести существенную ошибку, заменены значениями с учетом статистического распределения данных с столбе или строке в которой имеются отсутствующие данные. Например, отсутствующее значение может быть заменено на среднее или медианное значение соответствующего признака.

- при наличии возможности (достаточности данных) для поиска значений для замещения отсутствующих данных могут быть использованы алгоритмы машинного обучения, например алгоритмы линейной или логистической регрессии. В этом случае, имеющиеся данные используются для прогноза отсутствующих данных и помощью алгоритмов машинной обработки. Т.о. может быть создана последовательность MLалгоритмов как способ повышения качества исходного набора данных.

Краткий анализ рассмотренных подходов к работе с отсутствующими данными позволяет в качестве основного подхода остановиться на подходе, учитывающем временную упорядоченность демографических данных, что с учетом других особенностей демографических данных позволяет преступить к работе с отсутствующими данными.

В качестве примера возьмём фрагмент данных с 1989 по 2008 годы, который охватывает временной интервал в 18 лет.

Таблица 3. Работа с отсутствующими демографическими данными
Table 3. Dealing with missing demographic data

Группы	1989		1994		1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
0-4	12032				6762	6583	6511	6399	6530	6660	6916	7066	7234	7433
5-9	11360		12032		9314	8604	7940	6941	6852	6762	6583	6511	6503	6638
10-14	10592		11360		12544	12212	11852	10406	9860	9314	8604	7940	7458	7056
15-19	9968		10592		11870	12081	12098	12801	12673	12544	12212	11852	11244	10485
20-24	9755		9968		10797	10879	11054	11466	11668	11870	12081	12098	12298	12457
25-29	12557		9755		10030	10228	10316	10613	10705	10797	10879	11054	11130	11358
30-34	12863		12557		9665	9416	9427	9836	9933	10030	10228	10316	10466	10537
35-39	11684		12863		12155	11641	10925	10216	9941	9665	9416	9427	9485	9705
40-44	7663		11684		11891	11906	12070	12546	12351	12155	11641	10925	10325	9800
45-49	7955		7663		10447	10576	10738	11606	11749	11891	11906	12070	12084	11929
50-54	9593		7955		6466	7737	8724	10071	10259	10447	10576	10738	10887	11037
55-59	8399		9593		6387	5213	4458	5347	5907	6466	7737	8724	9164	9501
60-64	8360		8399		7021	7567	7699	7983	7185	6387	5213	4458	4408	5014
65-69	4510		8360					6345	6683	7021	7567	7699	7572	6687
70 и более	9646							12469	12397	12325	12242	12358	12605	13111

Источник: авторы, 2018.

В таблице 3 представлены результаты применения подхода к работе с отсутствующими данными с учетом их временной упорядоченности.

Подход 1. Использование средних значений. Применение этого подхода проиллюстрируем на фрагменте отсутствующих данных 2003 г. Как было определено выше, демографические данные в нормальных условиях отличаются последовательность и небольшим разбросом изменения. Это позволяет для оценки значения признаков по возрастным группам относящимся к 2003 году применить метод замены отсутствующих данных средним значением данных предшествующего и последующих наблюдения. Такое положение дел учитывает небольшой временной интервал, 1 год до и 1 год после, позволяет применить использование средних величин.

Оценка признака (2003) = (Признак 2004 + Признак 2002)/2 (2)

Это имеет отношение ко всем возрастным группам.

Результаты использования метода средних значений представлены в колонке 2003 и выделено красным цветом.

Подход 2. Метод перспективных оценок. Метод базируется на особенностях демографических данных, которые выявлены при построении «линий поколения». Его использование представляется целесообразным в том случае когда временные интервалы отсутствующих данных превышают 2-3 года. В нашем примере это соответствует интервалу в 5 лет, включающему отсутствующие данные за период 1990 – 2001 годы.

Перспективная оценка это перенос значения одной возрастной группы для использования в качестве оценки отсутствующих данных другой группы с учетом временного фактора. Например, состав возрастной группы «0-4» является базовой оценкой возрастной группы «5-9» с временным сдвигом вперед на 5 лет. А именно, численность группа «0-4 -1989» формирует основу оценки группы «5-9 1994».

«5-9 1994» ~ «0-4 1989» (3)

Результат применения метода перспективных оценок представлен в таблице 3 в колонке 1994 и выделено синим цветом. В отличии от метода средних оценок не все оценки по возрастным группам могут быть получены методом перспективных оценок. Так, например, данный метод не позволяет получать оценки по группе «0-4 1994», а также по возрастной группе «70 и более 1994» поскольку в этой группе сосредоточены все представители населения России в возрасте 70 лет и более и к этой возрастной группе не применимы закономерности свойственные другим возрастным группам.

Для оценки значения групп «0-4 1994» и «70 и более» следует применить более сложные вычисления с привлечением данных о численности населения «Всего»[14] и данных о рождаемости по годам с последующей группировкой по возрастным группам[15]. Для уточнения перспективной оценки можно предложить использование Коэффициентов выбытия, которые рассчитываются на основе имеющихся репрезентативных данных и данных государственной статистики о смертности по годам[16] с последующей группировкой по возрастным группам.

Подход 3. Метод ретроспективных оценок. Данный метод является обратным по отношению к методу перспективных оценок. Он основан на ретроспективном формировании оценки признака в текущем периоде возрастной группы путем присвоения ей значения которое берется у старшей возрастной группы за период 5 лет. Например, для оценки значения возрастной группы «0-4 2001» можно взять значение группы «5-9 2006» (См. табл 3. Выделено коричневым цветом). Аналогичным образом следует поступить для замещения отсутствующих данных за 1999 и 2000 годы. Однако, отсутствующие данные для возрастной группы «65-69» не могут быть оценены поскольку метод ретроспективных оценок не может быть применен к группе «70 и более» по причине описанной выше.

Временная упорядоченность демографических данных открывает простор для определения оценок оставшихся отсутствующих признаков путем составления и решения уравнений регрессии относительно неизвестных значений возрастных групп. Применение алгоритмов машинного обучения нельзя признать обоснованным в связи с тем, что обучающая выборка по частичным демографическим данным будет иметь серьезные недостатки с точки зрения достаточности объема и, как следствие, высокой вероятности получения некачественных оценок. К тому же объем данных для составления уравнения регрессии позволяет в полной мере решать задачу средствами Microsoft Exel.

Проектирование признаков (feature engineering). Эта процедура представляет собой процесс преобразования имеющихся исходных данных в подходящие для предсказывающей модели признаки. Состав проектируемых признаков определяется целями проводимого исследования и направлен на повышение качества исходного набора данных, способствующего более полному выявлению присущих данному набору закономерностей и решения машинными способами поставленной задачи.

В ходе исследования были разработан ряд признаков, которые могут быть использованы для повышения качества данных. Например, Коэффициент выбытия является одним из таких проектируемых признаков, раскрывающий определенные взаимосвязи между признаками.

Линии поколения не являются в полной мере проектируемыми признаками поскольку для их построения были применены только методы группировки и перегруппировки подходящих для решения этой задачи признаков.

Заключение

Проведенное исследования позволило получить следующие результаты:

- установлено, что демографические данные обладают рядом особенностей, которые могут быть и должны использоваться для повышения качества наборов данных для последующей обработки их с использованием технологий искусственного интеллекта и машинного обучения.

- во-первых, к особенностям демографических данных относится, в первую очередь, их временная упорядоченность.

- во-вторых, особенностью демографических данных является последовательность их изменения и отсутствие значительных перепадов между ближайшими значениями наблюдаемых данных.

- на демографические данные оказывают влияние различные процессы социально-политического и экономического характера в различные исторические периоды, что необходимо учитывать при работе с демографическими данными.

- демографическим данным, которые отнесятся к определенным историческим периодам, следует уделять особое внимание поскольку их значения могут как повышать эффективность набора данных для машинной обработки, так и быть причиной возникновения и роста систематических и случайных ошибок.

Для повышения качества данных и устранения недостатков, связанных с большим количеством отсутствующих значений в демографических данных, разработаны методы работы с отсутствующими данными к которым относятся:

- метод использования средних значений;
- метод перспективных оценок;
- метод ретроспективных оценок.

14 Федеральная служба государственной статистики. Интерактивная витрина [Электронный ресурс]. URL: http://cbsd.gks.ru/# (дата обращения: 23.09.2018).

15 Число родившихся за год. Единая межведомственная информационно-статистическая система (ЕМИСС) [Электронный ресурс]. URL: https://www.fedstat.ru/indicator/31606 (дата обращения: 23.09.2018).

16 Федеральная служба государственной статистики. Интерактивная витрина [Электронный ресурс]. URL: http://cbsd.gks.ru/# (дата обращения: 23.09.2018).

Представленные методы позволяют повысить эффективность работы с отсутствующими данными при наличии значительных пропусков. При необходимости оценки, полученные в результате применения указанных методов, могут уточняться путем корректировок на основе данных из вспомогательных источников или проектируемых признаков.

Результатом использования предлагаемых методов является формирование качественных данных с точки зрения их полноты и репрезентативности, что важно для их последующего анализа с использованием современных технологий на базе искусственного интеллекта и машинного обучения.

Список использованных источников

[1] *Соловьев В.И.* Анализ данных в экономике. Теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel. М.: КНОРУС, 2018. 479 с.

[2] *Бринк Х., Ричардс Д., Феверолф М.* Машинное обучение. СПб.: Питер, 2017. 336 с.

[3] *Bughin J.* et al. Artificial Intelligence: The Next Digital Frontier? Discussion Paper. McKinsey & Company, 2017. 78 p. URL: https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx (дата обращения: 23.09.2018).

[4] *Лушников А.А., Каган А.И., Гвишиани А.Д., Любовцева Ю.С.* Моделирование эволюционно-демографических процессов для целей геомедицины // Геофизические процессы и биосфера. 2013. Т. 12, № 3. С. 5-18. URL: https://elibrary.ru/item.asp?id=20265158 (дата обращения: 23.09.2018).

[5] *Поклонова Е.В., Захаренко П.В.* Статистический анализ динамики структуры безработного населения в РФ // Проблемы современной экономики (Новосибирск). 2014. № 22-2. С. 164-171. URL: https://elibrary.ru/item.asp?id=22676948 (дата обращения: 23.09.2018).

[6] *Божко Т.Н.* Демографическая оценка динамики численности и состава населения за 2010-2016 годы // Актуальные проблемы развития хозяйствующих субъектов, территорий и систем регионального и муниципального управления. 2017. С. 9-13. URL: https://elibrary.ru/item.asp?id=29371066 (дата обращения: 23.09.2018).

[7] *Карамнова Л.В., Копорова М.А.* Перспективы создания искусственного интеллекта // Гагаринские чтения 2017. М.: МАИ, 2017. С. 1178-1179. URL: https://elibrary.ru/item.asp?id=30084898 (дата обращения: 23.09.2018).

[8] *Дараган А.Д., Ежова Г.Л., Ежов Г.А.* О методологических аспектах реализации процедур обучения систем искусственного интеллекта // Современное педагогическое образование. 2017. № 3. С. 23-24. URL: https://elibrary.ru/item.asp?id=30731184 (дата обращения: 23.09.2018).

[9] *Андреев Е.М., Вишневский А.Г.* Ближайшие демографические перспективы России // Демоскоп Weekly. 2014. № 601-602. С. 1-25. URL: http://www.demoscope.ru/weekly/2014/0601/demoscope601.pdf (дата обращения: 23.09.2018).

[10] *Москвитин А.А., Созиев Т.М.* Особенности современных методов интеллектуального анализа данных // Современные методы интеллектуального анализа данных в экономических, гуманитарных и естественнонаучных исследованиях. Пятигорск, 2016. С. 11-18. URL: https://elibrary.ru/item.asp?id=30088652 (дата обращения: 23.09.2018).

[11] *Кривоносов Н.А.* Применение искусственного интеллекта как средства анализа/оценки деятельности логистических компаний // Шаг в будущее: искусственный интеллект и цифровая экономика. Вып. 3. М.: ГУУ, 2017. С. 99-106. URL: https://elibrary.ru/item.asp?id=32772302 (дата обращения: 23.09.2018).

[12] *Пидяшова О.П.* Исследования уровня жизни населения в современных условиях (Региональный аспект) // Экономика и предпринимательство. 2015. № 12-1(65). С. 1083-1090. URL: https://elibrary.ru/item.asp?id=25031618 (дата обращения: 23.09.2018).

[13] *Соловьев А.И.* Человек в цифровой экономике: аналоговый или дискретный? // Научные труды Вольного экономического общества России. 2018. Т. 210, № 2. С. 130-134. URL: https://elibrary.ru/item.asp?id=35041906 (дата обращения: 23.09.2018).

[14] *Куприяновский В.П., Сотников А.Е., Соловьев А.И., Дрожжинов В.И., Намиот Д.Е., Мамаев В.Ю., Куприяновский П.В.* AADHAAR – идентификация человека в цифровой экономике // International Journal of Open Information Technologies. 2017. Т. 5, № 2. С. 34-45. URL: https://elibrary.ru/item.asp?id=28314924 (дата обращения: 23.09.2018).

[15] *Gepp A., Linnenluecke M.K., O'Neill T.J.* Tom Smith Big data techniques in auditing research and practice: Current trends and future opportunities // Journal of Accounting Literature. 2018. Vol. 40. Pp. 102-115. DIO: 10.1016/j.acclit.2017.05.003

[16] *Kauffman R.J., Kim K., Lee S-Y.T. Hoang A-P., Ren J.* Combining machine-based and econometrics methods for policy analytics insights // Electronic Commerce Research and Applications. 2017. Vol. 25, issue C. Pp. 115-140. DOI: 10.1016/j.elerap.2017.04.004

[17] *Dremel C., Herterich M.M., Wulf J., vom Brocke J.* Actualizing Big Data Analytics Affordances: A Revelatory Case Study // Information & Management. 2018. DOI: 10.1016/j.im.2018.10.007

[18] *McClean S.I.* Data Mining and Knowledge Discovery // Encyclopedia of Physical Science and Technology. (Third Edition), 2003. Pp. 229-246. DOI: 10.1016/B0-12-227410-5/00845-0

Поступила 23.09.2018; принята в печать 10.11.2018;
опубликована онлайн 10.12.2018.

References

[1] Soloviev V.I. Analysis of data in the economy. Probability theory, applied statistics, data processing and visualization in Microsoft Excel. M.: KNORUS, 2018. 479 p. (In Russian)

[2] Brink H., Richards J., Fetherolf M. Real-World Machine Learning. Manning, 2016. 264 p.

[3] Bughin J. et al. Artificial Intelligence: The Next Digital Frontier? Discussion Paper. McKinsey & Company, 2017. 78 p. Available at: https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx (accessed

23.09.2018).
[4] Lushnikov A.A., Kagan, A.I., Gvishiani, A.D., Lyubovtseva, Yu.S. Modeling of the evolutionary demographic processes for geomedicine. *Geophysical Pprocesses and Biosphere*. 2013; 12(3):5-18. Available at: https://elibrary.ru/item.asp?id=20265158 (accessed 23.09.2018). (In Russian)
[5] Poklonova E.V., Zakharenko P.V. Statistical analysis of the dynamics of the structure of the unemployed population in the Russian Federation. *Problems of the Modern Economy (Novosibirsk)*. 2014; 22-2:164-171. Available at: https://elibrary.ru/item.asp?id=22676948 (accessed 23.09.2018). (In Russian)
[6] Bozhko T.N. Demographic assessment of the dynamics of the number and composition of the population for 2010-2016. *Proceedings of the Actual problems of the development of economic entities, territories and systems of regional and municipal government.* 2017, pp. 9-13. Available at: https://elibrary.ru/item.asp?id=29371066 (accessed 23.09.2018). (In Russian)
[7] Karamnova L.V., Koporova M.A. Prospects for creating artificial intelligence. *Gagarin Readings* 2017. M.: MAI, 2017, pp. 1178-1179. Available at: https://elibrary.ru/item.asp?id=30084898 (accessed 23.09.2018). (In Russian)
[8] Daragan A.D., Ezhov G.L., Ezhov G.A. On the methodological aspects of the implementation of training procedures for artificial intelligence systems. *Modern Pedagogical Education*. 2017; 3:23-24. Available at: https://elibrary.ru/item.asp?id=30731184 (accessed 23.09.2018). (In Russian)
[9] Andreev E.M., Vishnevsky A.G. The nearest demographic perspectives of Russia. *Demoscope Weekly*. 2014; 601-602:1-25. Available at: http://www.demoscope.ru/weekly/2014/0601/demoscope601.pdf (accessed 23.09.2018). (In Russian)
[10] Moskvitin A.A., Soziev T.M. Features of modern methods of data mining. *Modern Methods of Data Mining in Economic, Humanitarian, and Natural Sciences. Proceedings of the international scientific-practical conference*. Pyatigorsk, 2016. pp. 11-18. Available at: https://elibrary.ru/item.asp?id=30088652 (accessed 23.09.2018). (In Russian)
[11] Krivonosov N.A. The use of artificial intelligence as a means of analyzing / evaluating the activities of logistics companies. *A Step to the Future: Artificial Intelligence and the Digital Economy. Proceedings of the 1st International Scientific Practical Conference*. Vol. 3. M.: SUM, 2017, pp. 99-106. Available at: https://elibrary.ru/item.asp?id=32772302 (accessed 23.09.2018). (In Russian)
[12] Pidyashova O.P. Studies of the standard of living of the population in modern conditions (regional aspect). *Journal of Economy and entrepreneurship*. 2015; 12-1:1083-1090. Available at: https://elibrary.ru/item.asp?id=25031618 (accessed 23.09.2018). (In Russian)
[13] Solovev A.I. A Person in a Digital Economy: Analog or Discrete? *Scientific Works of the Free Economic Society of Russia*. 2018; 210(2):130-134. Available at: https://elibrary.ru/item.asp?id=35041906 (accessed 23.09.2018). (In Russian)
[14] Kupriyanovskiy V.P., Sotnikov A.E., Solovev A.I., Drozhzhinov V.I., Namiot D.E., Mamaev V.Yu., Kupriyanovskiy P.V. AADHAAR - Identification of the Person in the Digital Economy. *International Journal of Open Information Technologies*. 2017; 5(2):34-45. Available at: https://elibrary.ru/item.asp?id=28314924 (accessed 23.09.2018). (In Russian)
[15] Gepp A., Linnenluecke M.K., O'Neill T.J. Tom Smith Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature*. 2018; 40:102-115. DIO: 10.1016/j.acclit.2017.05.003
[16] Kauffman R.J., Kim K., Lee S-Y.T. Hoang A-P., Ren J. Combining machine-based and econometrics methods for policy analytics insights. *Electronic Commerce Research and Applications*. 2017; 25(C):115-140. DOI: 10.1016/j.elerap.2017.04.004
[17] Dremel C., Herterich M.M., Wulf J., vom Brocke J. Actualizing Big Data Analytics Affordances: A Revelatory Case Study. *Information & Management*. 2018. DOI: 10.1016/j.im.2018.10.007
[18] McClean S.I. Data Mining and Knowledge Discovery. *Encyclopedia of Physical Science and Technology.* (Third Edition), 2003, pp. 229-246. DOI: 10.1016/B0-12-227410-5/00845-0

Submitted 23.09.2018; revised 10.11.2018;
published online 10.12.2018.

About the authors:

Anatolii I. Solovev, Candidate of Technical Sciences, Associate Professor, Department of Data Analysis, Decision Making and Financial Technologies, Financial University under the Government of the Russian Federation (49 Leningradsky Av., Moscow 125993, GSP-3, Russia), ORCID: http://orcid.org/0000-0002-8544-7973, aisol@rambler.ru
Stefan A. Solovev, Postgraduate Student of the Corporate Finance and Corporate Governance Department, Financial University under the Government of the Russian Federation (49 Leningradsky Av., Moscow 125993, GSP-3, Russia), ORCID: http://orcid.org/0000-0001-9443-1423, StefanSolovyev@me.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted reuse, distribution, and reproduction in any medium provided the original work is properly cited.