Вычислительные технологии, 2020, том 25, № 5, с. 80–90. © ФИЦ ИВТ, 2020
Computational Technologies, 2020, vol. 25, no. 5, pp. 80–90. © FRC ICT, 2020
ISSN 1560-7534
eISSN 2313-691X

DOI:10.25743/ICT.2020.25.5.007

Применение методов искусственного интеллекта и сжатия данных для прогнозирования социальных, экономических и демографических показателей Новосибирской области

К. С. Чирихин[1,2,*], Б. Я. Рябко[1,2]

[1]Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск, Россия

[2]Новосибирский государственный университет, Новосибирск, Россия

[*]Контактный автор: Чирихин Константин Сергеевич, e-mail: chirihin@gmail.com

Поступила 14 апреля 2020 г., доработана 15 сентября 2020 г., принята в печать 21 сентября 2020 г.

Рассмотрена задача прогнозирования временны́х рядов в ее одно- и многомерной постановках. Описан метод прогнозирования, в основе которого лежат алгоритмы сжатия данных и искусственного интеллекта. В качестве примера практического применения этого метода приведены вычисления для некоторых социальных, демографических и экономических показателей Новосибирской области. Наряду с широко используемыми на практике программами для сжатия данных, использованы реализации относительно менее известных моделей на основе формальных грамматик, а также метода, основанного на конечных автоматах. Для прогнозных значений построены доверительные интервалы.

Ключевые слова: универсальное кодирование, многомерные временны́е ряды, искусственный интеллект.

Цитирование: Чирихин К.С., Рябко Б.Я. Применение методов искусственного интеллекта и сжатия данных для прогнозирования социальных, экономических и демографических показателей Новосибирской области. Вычислительные технологии. 2020; 25(5):80–90. DOI:10.25743/ICT.2020.25.5.007.

Введение

Задача прогнозирования временны́х рядов обладает большой практической значимостью и находит применение во многих областях человеческой деятельности. В одномерном случае она заключается в экстраполяции последовательности числовых значений, упорядоченных во времени, на несколько значений вперед. Существует и многомерная (или векторная) постановка этой задачи, в которой требуется совместно прогнозировать несколько взаимосвязанных последовательностей. Учет зависимостей между рядами может привести к повышению точности прогноза. Например, временны́е ряды температуры воздуха, атмосферного давления, скорости и направления ветра связаны между собой, и их совместное прогнозирование может иметь смысл. Задача многомерного прогнозирования нашла применение в науках о земле, экономике, инженерных приложениях [1].

К настоящему времени предложены разнообразные подходы к прогнозированию временны́х рядов, среди которых отметим модели авторегрессии-скользящего среднего (ARMA), обобщенной авторегрессионной условной гетероскедастичности (GARPH) (их описание может быть найдено, например, в [2]), нейронные сети [3], экспоненциальное сглаживание [4]. Многие методы имеют обобщения на многомерный случай. Например, векторная модель ARMA носит название VARMA [2]. Модель GARPH также может быть использована для прогнозирования многомерных временны́х рядов [5]. Тем не менее лучшего универсального метода не существует и задача построения новых методов прогнозирования по-прежнему остается актуальной.

В данной работе развит подход, подробно изложенный в [6]. С использованием набора программ для сжатия данных выполнено прогнозирование одно- и многомерных временны́х рядов некоторых демографических, социальных и экономических показателей Новосибирской области. В этот набор программ включены реализации широко используемых на практике алгоритмов сжатия, реализации относительно менее известных алгоритмов, основанных на формальных грамматиках, а также наша реализация алгоритма на основе детерминированных конечных автоматов с несколькими головками. Для прогнозных значений построены доверительные интервалы.

1. Прогнозирование одномерных временны́х рядов

Первоначально подход к прогнозированию временны́х рядов с использованием методов сжатия данных предложен в [7]. Приведем его основную идею. Рассмотрим временной ряд x_1, x_2, ..., x_t, в котором x_i принадлежат некоторому конечному алфавиту (множеству) A. Представим прогноз в виде условных вероятностей $p(x_{t+1} = a|x_1, x_2, \ldots, x_t)$, $x_i, a \in A$. Их оценки p^* могут быть получены при помощи произвольного метода сжатия данных ϕ, позволяющего однозначно восстановить исходное сообщение без потерь, по следующей формуле:

$$p^*(x_{t+1} = a|x_1, x_2, \ldots, x_t) = \frac{2^{-|\phi(x_1,x_2,\ldots,x_t,a)|}}{\sum\limits_{b\in A} 2^{-|\phi(x_1,x_2,\ldots,x_t,b)|}},$$

где $|\phi(\alpha)|$ — размер (в битах) сжатого представления последовательности α при кодировании с помощью метода ϕ. В общем случае, при прогнозировании на h шагов вперед, $a, b \in A^h$, где A^h — множество всех последовательностей длины h над алфавитом A.

Несколько алгоритмов сжатия можно объединить в один метод прогнозирования:

$$p^*(x_{t+1} = a|x_1, x_2, \ldots, x_t) = \frac{\sum\limits_{i=1}^{m} \omega_i 2^{-|\phi_i(x_1,x_2,\ldots,x_t,a)|}}{\sum\limits_{b\in A}\sum\limits_{i=1}^{m} \omega_i 2^{-|\phi_i(x_1,x_2,\ldots,x_t,b)|}}. \quad (1)$$

Здесь m — количество объединяемых алгоритмов; ω_i — неотрицательные весовые коэффициенты; $\sum\limits_{i=1}^{m} \omega_i = 1$. Асимптотически выбор ω_i не влияет на эффективность алгоритма.

Поскольку при сжатии данных устраняются найденные в них закономерности, алгоритм, лучше остальных сжимающий ряд, наиболее подходит для его прогнозирования. Именно он оказывает основное влияние на результат в (1), при этом вклад других алгоритмов практически не заметен. Таким образом, (1) как бы “выбирает” наилучший

алгоритм для прогнозирования ряда. Объединяя алгоритмы, эффективные для поиска закономерностей разных типов, получаем универсальный метод. Очевидный недостаток подобного подхода — его трудоемкость, поскольку возникающие при прогнозировании последовательности должны быть сжаты всеми алгоритмами (а количество таких последовательностей увеличивается экспоненциально с ростом числа шагов, на которые строится прогноз).

Добавим, что методы сжатия данных могут использоваться для прогнозирования вещественных временны́х рядов. В этом случае ряд предварительно преобразуется к ряду с конечным алфавитом с помощью процедуры квантования, которая в простейшем варианте заключается в разбиении отрезка, содержащего все значения ряда, на некоторое конечное число равных пронумерованных интервалов и последующей замене каждого элемента ряда номером интервала, в который попадает этот элемент. Существует способ объединения прогнозов, полученных с использованием разбиений указанного отрезка на разное количество интервалов (его описание может быть найдено, например, в [8]). Он заключается в совместном рассмотрении нескольких разбиений, количество интервалов в каждом из которых является степенью 2, и "выборе" наиболее подходящего разбиения с помощью подхода, аналогичного (1).

2. Прогнозирование многомерных временны́х рядов

Обобщим изложенный ранее метод на многомерный случай. Для простоты рассмотрим только временны́е ряды с конечными алфавитами, поскольку обобщение на случай непрерывного алфавита для многомерных рядов аналогично одномерным. Предположим, что есть k временных рядов X_1, X_2, ..., X_k, $X_j = x_{1j}, x_{2j}, \ldots, x_{tj}$ и их элементы принимают значения из алфавита $A = \{0, 1, 2, \ldots, n-1\}$. Пусть мы хотим построить прогноз на $h \geq 1$ шагов вперед для каждого ряда. Можно свести эту задачу к одномерному случаю, построив новый ряд $X' = x'_1, x'_2, \ldots, x'_t$ по следующему правилу: $x'_i = \sum_{j=1}^{k} x_{ij}|A|^{j-1}$. Затем по прогнозу для X' легко получить прогноз для каждого из k исходных рядов.

Пример 1. Рассмотрим преобразование рядов

$$X_1 = 3208, -355, 2163, 2807, -154, 1760, -2234, -3706,$$

$$X_2 = 6385, -5729, 1279, 1924, -2513, 1173, 1442, -2837$$

к одномерному ряду целых чисел, который можно непосредственно прогнозировать с помощью методов сжатия данных. Несмотря на то что X_1 и X_2 являются целочисленными, применим процедуру квантования для уменьшения размера алфавита. Наименьший элемент первого ряда равен -3706, наибольший — 3208. Для второго ряда это значения -5729 и 6385 соответственно. Для простоты будем рассматривать только разбиения на два интервала: $A = \{0, 1\}$. Соответствие между получающимися интервалами и их номерами приведено в табл. 1.

Выполнив замену, получим следующие два ряда:

$$X_1 = 1, 0, 1, 1, 1, 1, 0, 0,$$

$$X_2 = 1, 0, 1, 1, 0, 1, 1, 0.$$

Т а б л и ц а 1. Соответствие между интервалами и их номерами при квантовании для рядов из примера 1
Table 1. Correspondence between intervals and their numbers for quantizing time series from example 1

Ряд	Номер интервала	
	0	1
X_1	$[-3706; -249)$	$[-249; 3208]$
X_2	$[-5729; 328)$	$[328; 6385]$

Применив описанную процедуру перехода от многомерного случая к одномерному, получим следующий временной ряд:

$$X' = x'_1, x'_2, \ldots, x'_8 = 3, 0, 3, 3, 1, 3, 2, 0.$$

Подробно рассмотренный пример прогнозирования одномерного временно́го ряда с помощью методов сжатия данных может быть найден в [8].

3. Прогнозирование экономических, социальных и демографических показателей Новосибирской области

Применим описанный метод для прогнозирования некоторых демографических, социальных и экономических показателей Новосибирской области (НСО). Временны́е ряды, которые рассматриваем, могут быть найдены на официальном сайте Федеральной службы государственной статистики по НСО (`https://novosibstat.gks.ru`, дата обращения 15.03.2020). Программная реализация метода, с помощью которой получены прогнозные значения, доступна по адресу `https://github.com/kchirikhin/itp`. Опишем использованную нами методику вычислений, а затем приведем полученные прогнозные значения.

Для построения прогнозов использованы следующие 5 программ, в которых реализованы разные подходы к сжатию данных:

- *zlib* — библиотека для сжатия данных, в которой реализована схема DEFLATE. Используется в широко известной программе gzip. Адрес сайта проекта `https://zlib.net`, в настоящей работе использована версия 1.2.11;
- *bzip2* — программа для сжатия данных, в которой реализованы алгоритм Барроуза — Уилера и код "стопка книг" (также известный под названием move-to-front). Доступна по адресу `https://www.sourceware.org/bzip2`;
- *ppmd* — вариант алгоритма предсказания по частичному совпадению (prediction by partial matching, PPM). В работе использована реализация, исходный код которой может быть найден по адресу `https://github.com/Shelwien/ppmd_sh`;
- *rp* — алгоритм сжатия данных на основе контекстно-свободных грамматик. Использована реализация, которая доступна по адресу `https://github.com/nicolaprezza/Re-Pair`;
- *automaton* — алгоритм на основе детерминированного конечного автомата с десятью головками. Он способен правильно прогнозировать слова с закономерностями вида 01001000100001... (так называемые полилинейные слова). Алгоритм для бесконечных полилинейных слов предложен в [9], его модификация для прогнозирования временны́х рядов и использования совместно с методами сжатия приведена в [10]. Мы самостоятельно реализовывали данный алгоритм.

Поскольку при прогнозировании с помощью методов сжатия прогноз представлен в виде распределения вероятностей, в качестве прогнозных значений брались математические ожидания, вычисленные по полученным распределениям. При объединении распределений от разных алгоритмов с помощью (1) использованы равные веса, т. е. априорно не отдано предпочтение какому-либо алгоритму. Для представления одного элемента временно́го ряда использовался 1 байт.

Чтобы удалить присутствующие в прогнозируемых рядах тренды, брали от них первую разность — вместо ряда $x_1, x_2, \ldots, x_t$ прогнозировали ряд $x_2{-}x_1, x_3{-}x_2, \ldots, x_t{-}x_{t-1}$. Для некоторых рядов брали вторую разность (в таких случаях это явно указано).

Предположим, что рассматриваем ряд $X = x_1, x_2, \ldots, x_t$ и вычисляем прогноз на h шагов вперед. Чтобы построить доверительные интервалы для прогнозных значений $\hat{x}_{t+i}$, $1 \le i \le h$, оценивали стандартные отклонения ошибок прогнозов уже известных значений. Более точно, начиная с элемента с номером $j = \lfloor t/2 \rfloor + 1$, перед добавлением x_j в конец ряда $x_1, x_2, \ldots, x_{j-1}$ строили прогноз для следующих h значений и вычисляли ошибки прогнозирования как $|x_{j+i} - \hat{x}_{j+i}|$, где x_{j+i} — зафиксированное значение, $\hat{x}_{j+i}$ — прогнозное значение. По совокупности таких ошибок для каждого шага вычисляли среднее квадратическое отклонение σ_i. В качестве доверительного интервала для прогнозного значения брали интервал $\hat{x}_{t+i} \pm 2\sigma_i$, что приблизительно соответствует доверительной вероятности 0.954 для нормального распределения.

Описание результатов вычислений начнем со случаев, в которых применение многомерного прогнозирования привело к повышению точности прогнозов. Рассмотрим ряды среднегодовой численности и естественного прироста населения в Новосибирской области (НСО). Их графики представлены на рис. 1. Здесь же приведены прогнозы для каждого ряда на три года вперед, полученные с помощью многомерного прогнозирования. Для их построения применялась процедура взятия второй разности, а также совместно использовались разбиения отрезка, в который попадают все значения соответствующего ряда, на 2, 4 и 8 интервалов. Отметим, что на указанный момент обращения к сайту Федеральной службы государственной статистики по НСО последние опубликованные значения для данных показателей были за 2018 г. В табл. 2 приведены прогнозные значения с доверительными интервалами, полученные при много- и одномерном прогнозировании, а в табл. 3 — средние относительные ошибки для каждого шага (1–3), вычисленные по ошибкам при прогнозировании уже зафиксированных значений (начиная с 2011 г.). Как видно из этой таблицы, средняя относительная ошибка при многомерном прогнозировании оказалась ниже.

В качестве другого примера рассмотрим прогнозирование рядов ожидаемой продолжительности жизни и количества умерших в НСО. Их графики, вместе с прогнозными значениями на три шага вперед и доверительными интервалами, приведены на рис. 2. Мы брали первую разность, в остальном вычисления проводились аналогично предыдущему случаю. В табл. 4 полученные прогнозные значения и интервалы представлены в числовой форме. Сравнение средних относительных ошибок при одно- и многомерном прогнозировании приведено в табл. 5. Как видно из этой таблицы, средняя относительная ошибка для количества умерших в НСО оказалась ниже при многомерном прогнозировании, в то время как для ожидаемой продолжительности жизни однозначный вывод сделать нельзя.

В табл. 6 приведены прогнозы на четыре года вперед для некоторых других показателей НСО, полученные с помощью одномерного прогнозирования. Увеличение количества шагов по сравнению с предыдущими примерами связано с тем, что трудоемкость

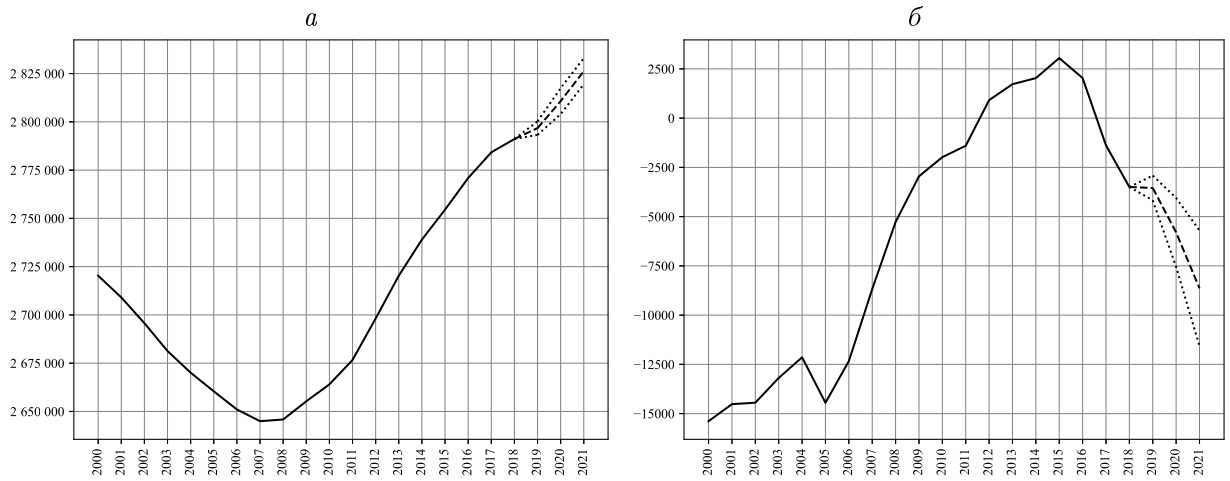

Рис. 1. Временны́е ряды среднегодовой численности населения (*а*) и естественного его прироста (*б*) в НСО: непрерывные линии — зафиксированные значения, штриховые — прогнозные, пунктирными линиями ограничены доверительные интервалы (уровень доверия 95 %)
Fig. 1. Average annual population (*а*) and natural population growth (*б*) in the Novosibirsk region. Solid lines show observed values, dashed lines indicate predicted values, and dotted lines bound 95 % confidence intervals

Т а б л и ц а 2. Прогнозные значения, полученные путем одно- и многомерного прогнозирования рядов среднегодовой численности и естественного прироста населения в НСО
Table 2. Univariate and multivariate forecasts of the average annual population and natural population growth in the Novosibirsk region

Тип прогнозирования	Ряд	2019	2020	2021
Одномерное	Среднегодовая численность населения	2 795 721 [2 791 892; 2 799 550]	2 807 615 [2 800 208; 2 815 021]	2 820 005 [2 811 159; 2 828 851]
	Естественный прирост населения	−2969 [−3557; −2381]	−5296 [−7153; −3439]	−8572 [−11764; −5381]
Многомерное	Среднегодовая численность населения	2 796 757 [2 793 253; 2 800 261]	2 810 730 [2 804 002; 2 817 458]	2 826 252 [2 819 460; 2 833 043]
	Естественный прирост населения	−3545 [−4171; −2919]	−5801 [−7551; −4051]	−8637 [−11571; −5702]

Т а б л и ц а 3. Сравнение средних относительных ошибок, полученных при одно- и многомерном прогнозировании рядов среднегодовой численности и естественного прироста населения в НСО
Table 3. Mean relative errors of the univariate and multivariate forecasts of the average annual population and natural population growth in the Novosibirsk region

Тип прогнозирования	Ряд	Шаг 1	Шаг 2	Шаг 3
Одномерное	Среднегодовая численность населения	0.004	0.005	0.007
	Естественный прирост населения	0.143	0.356	0.670
Многомерное	Среднегодовая численность населения	0.004	0.003	0.006
	Естественный прирост населения	0.138	0.278	0.545

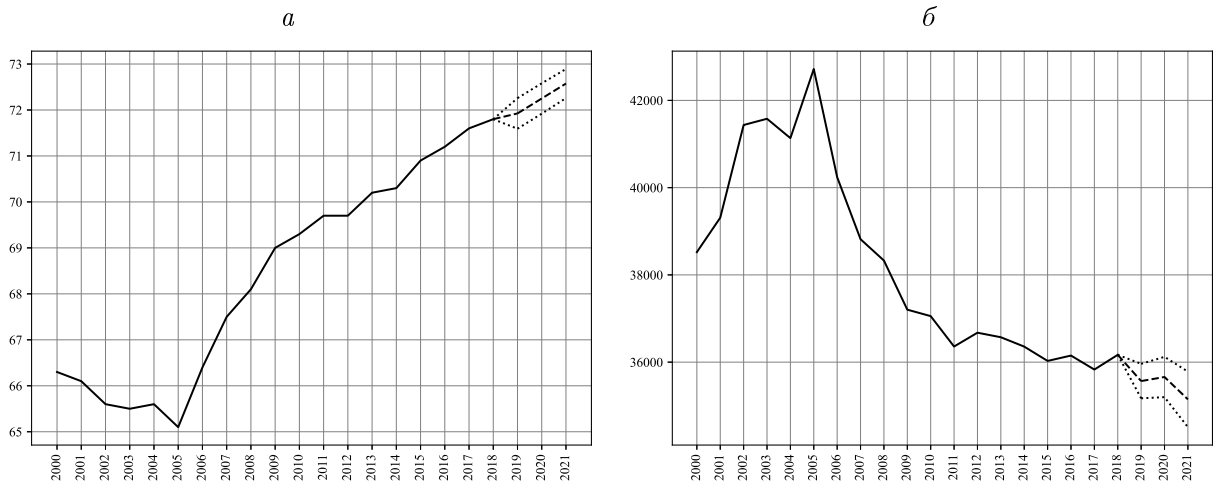

Рис. 2. Временны́е ряды ожидаемой продолжительности жизни (*а*) и количества умерших (*б*) в НСО: непрерывные линии — зафиксированные значения, штриховые — прогнозные, пунктирными линиями ограничены доверительные интервалы (уровень доверия 95 %)
Fig. 2. Life expectancy (*а*) and number of deaths (*б*) in the Novosibirsk region. Solid lines show observed values, dashed lines indicate predicted values, and dotted lines bound 95 % confidence intervals

Т а б л и ц а 4. Прогнозные значения, полученные путем одно- и многомерного прогнозирования рядов ожидаемой продолжительности жизни и количества умерших в НСО
Table 4. Univariate and multivariate forecasts of the life expectancy and number of deaths in the Novosibirsk region

Тип прогнозирования	Ряд	2019	2020	2021
Одномерное	Ожидаемая продолжительность жизни	72.27 [72.05; 72.49]	72.25 [72.05; 72.45]	72.92 [72.67; 73.17]
	Количество умерших	35 846 [35 478; 36 214]	35 740 [35 276; 36 205]	35 438 [34 897; 35 978]
Многомерное	Ожидаемая продолжительность жизни	71.93 [71.59; 72.26]	72.25 [71.92; 72.58]	72.58 [72.26; 72.89]
	Количество умерших	35 566 [35 171; 35 961]	35 660 [35 198; 36 121]	35 147 [34 510; 35 784]

Т а б л и ц а 5. Сравнение средних относительных ошибок, полученных при одно- и многомерном прогнозировании рядов ожидаемой продолжительности жизни и количества умерших в НСО
Table 5. Mean relative errors of the univariate and multivariate forecasts of the life expectancy and number of deaths in the Novosibirsk region

Тип прогнозирования	Ряд	Шаг 1	Шаг 2	Шаг 3
Одномерное	Ожидаемая продолжительность жизни	0.005	0.005	0.009
	Количество умерших	0.014	0.020	0.036
Многомерное	Ожидаемая продолжительность жизни	0.006	0.006	0.006
	Количество умерших	0.013	0.015	0.017

Т а б л и ц а 6. Прогнозные значения на 4 шага вперед и доверительные интервалы для некоторых других показателей НСО

Table 6. 4-steps-ahead forecasts with confidence intervals for some other indicators of the Novosibirsk region

Показатель	Шаг 1	Шаг 2	Шаг 3	Шаг 4
Прожиточный минимум, руб.	2020	2021	2022	2023
	11 405 [10 777; 12 033]	11 752 [11 106; 12 398]	12 098 [11 394; 12 802]	12 445 [11 897; 12 992]
Валовой региональный продукт, млн руб.	2019	2020	2021	2022
	1 274 481 [1 243 953; 1 305 009]	1 380 703 [1 323 329; 1 438 077]	1 499 047 [1 451 156; 1 546 939]	1 627 328 [1 558 722; 1 695 934]
Количество браков	2019	2020	2021	2022
	17 190 [16 592; 17 787]	14 585 [13 482; 15 689]	11 803 [9786; 13 819]	8590 [5988; 11 191]
Количество разводов	2019	2020	2021	2022
	12 699 [11 721; 13 677]	11 904 [10 954; 12 855]	11 799 [10 631; 12 967]	11 001 [9355; 12 646]
Средний возраст матери	2019	2020	2021	2022
	28.94 [28.85; 29.03]	29.1 [28.96; 29.24]	29.27 [29.1; 29.43]	29.44 [29.29; 29.58]
Дебиторская задолженность организаций, млн руб.	2019	2020	2021	2022
	349 623 [345 045; 354 201]	372 745 [359 634; 385 856]	397 037 [394 762; 399 312]	418 228 [407 720; 428 737]
Кредиторская задолженность организаций, млн руб.	2019	2020	2021	2022
	458 611 [446 726; 470 496]	480 336 [473 660; 487 011]	492 491 [486 058; 498 924]	514 131 [472 756; 555 506]
Средняя цена на рынке жилья, руб. за 1 кв. м	2020	2021	2022	2023
	56 743 [53 363; 60 123]	60 273 [55 523; 65 022]	63 766 [58 357; 69 174]	66 952 [61 477; 72 428]

вычислений при одномерном прогнозировании ниже, чем при многомерном. Мы брали вторую разность при прогнозировании валового регионального продукта и количества браков, во всех остальных случаях использовалась первая разность. Как и ранее, совместно рассматривались разбиения множеств возможных значений рядов на 2, 4 и 8 интервалов при квантовании.

4. Анализ результатов вычислений

Согласно результатам проведенных вычислений, в ближайшее время среднегодовая численность населения в НСО будет увеличиваться и к 2021 г. превысит 2 млн 820 тыс. человек, при этом естественный прирост населения будет отрицательным. Ожидаемая продолжительность жизни также будет возрастать, а количество умерших останется приблизительно на текущем уровне. Валовой региональный продукт Новосибирской области продолжит увеличиваться и к 2022 г. превысит 1.6 трлн рублей. Прожиточный минимум к 2023 г. увеличится на 9.5 % по сравнению с 2019 г. Количество браков будет

сокращаться быстрее количества разводов. За ближайшие четыре года средний возраст матери повысится приблизительно на 0.5 года, дебиторская задолженность организаций вырастет на 19.6 %, а кредиторская — на 12 %. Средняя стоимость квадратного метра на рынке жилья увеличится на 10 тыс. рублей.

Заключение

Описан подход к прогнозированию временны́х рядов, основанный на методах сжатия данных и искусственного интеллекта, а также приведены прогнозные значения для основных показателей Новосибирской области, полученные с его помощью. Результаты вычислений, на наш взгляд, показывают, что описываемый метод способен находить нетривиальные закономерности в данных и может использоваться на практике. Отметим, что работа выполнена до наступления пандемии коронавируса и в ней не учтены ее влияния, так как события подобного рода не наступали в течение всего периода фиксации показателей Новосибирской области. Тем не менее полученные данные представляют интерес хотя бы для последующего сравнения прогнозных значений, сделанных до пандемии, с реальными, измеренными после ее окончания.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (гранты №№ 19-37-90009, 19-47-540001).

Список литературы

[1] **Reinsel G.C.** Elements of multivariate time series analysis. New York: Springer Science & Business Media; 2003: 380.

[2] **Box G.E., Jenkins G.M., Reinsel G.C., Ljung G.M.** Time series analysis: Forecasting and control. John Wiley & Sons; 2015: 712.

[3] **Kaastra I., Boyd M.** Designing a neural network for forecasting financial and economic time series. Neurocomputing. 1996; (10):215–236.

[4] **Winters P.R.** Forecasting sales by exponentially weighted moving averages. Management Science. 1960; 6(3):324–342.

[5] **Engle R.F., Kroner K.F.** Multivariate simultaneous generalized ARCH. Econometric Theory. 1995; 11(1):122–150.

[6] **Ryabko B., Astola J., Malyutov M.** Compression-based methods of statistical analysis and prediction of time series. Switzerland: Springer International Publishing; 2016: 144.

[7] **Рябко Б.Я.** Прогноз случайных последовательностей и универсальное кодирование. Проблемы передачи информации. 1988; 24(2):3–14.

[8] **Чирихин К.С., Рябко Б.Я.** Экспериментальное исследование точности методов прогноза, базирующихся на архиваторах. Вест. Новосиб. гос. ун-та. Сер. Информац. технол. 2018; 16(3):145–158.

[9] **Tim S.** Prediction of infinite words with automata. Theory of Computing Systems. 2018; (62):653–681.

[10] **Chirikhin K.S., Ryabko B.Ya.** Application of artificial intelligence and data compression methods to time series forecasting. Proceedings of the International Workshop “Applied Methods of Statistical Analysis. Statistical Computation and Simulation — AMSA’2019”. Novosibirsk. 2019: 553–560.

Вычислительные технологии, 2020, том 25, № 5, с. 80–90. © ФИЦ ИВТ, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 5, pp. 80–90. © FRC ICT, 2020 eISSN 2313-691X

COMPUTATIONAL TECHNOLOGIES

DOI:10.25743/ICT.2020.25.5.007

The application of artificial intelligence and data compression techniques for forecasting of social, economic and demographic indicators of the Novosibirsk region

Chirikhin Konstantin S.[1,2,*], Ryabko Boris Ya.[1,2]

[1]Federal Research Center for Information and Computational Technologies, Novosibirsk, Russia
[2]Novosibirsk State University, Novosibirsk, Russia

[*]Corresponding author: Chirikhin Konstantin S., e-mail: chirihin@gmail.com
Received April 14, 2020, revised September 15, 2020, accepted September 21, 2020

Abstract

This paper describes and experimentally investigates a time series forecasting method based on data compression and artificial intelligence techniques. Its basic idea is to combine various algorithms that can estimate the compressed size of a sequence of discrete values into a single method of forecasting. Generalizations of the method to continuous and multivariate cases are described. We use several popular data compression libraries (zlib, bzip2, ppmd), as well as applications of relatively lesser-known algorithms based on formal grammars (re-pair) along with our implementation of an algorithm based on finite automata with multiple heads. All these are employed to make forecasts for some social, economic and demographic indicators of the Novosibirsk region. The article elaborates both our methodology and the methods used for data preprocessing. Confidence intervals with a confidence level of 0.95 are plotted for all predicted values; the average relative errors calculated from the results of predicting the already fixed values are given. Cases in which multivariate forecasting turned out to be more accurate than univariate forecasting are:

1. Average annual population and natural population growth in the Novosibirsk region. Figure 1 shows their graphs along with multivariate predictions, while confidence intervals and mean relative errors are presented in tables 2, 3.
2. Life expectancy and number of deaths in the Novosibirsk region. Figure 2 shows their graphs along with multivariate predictions, confidence intervals and mean relative errors are presented in tables 4, 5. Table 6 gives univariate forecasts of some other indicators of the Novosibirsk region.

We think that the results of computations show that the proposed method is capable of finding non-trivial patterns in the data and can be used in practice.

Keywords: universal coding, multivariate time series, artificial intelligence.

Citation: Chirikhin K.S., Ryabko B.Ya. The application of artificial intelligence and data compression techniques for forecasting of social, economic and demographic indicators of the Novosibirsk region. Computational Technologies. 2020; 25(5):80–90. DOI:10.25743/ICT.2020.25.5.007. (In Russ.)

Acknowledgements. This research was partly supported by RFBR (grants No. 19-37-90009, 19-47-540001).

References

1. **Reinsel G.C.** Elements of multivariate time series analysis. New York: Springer Science & Business Media; 2003: 380.
2. **Box G.E., Jenkins G.M., Reinsel G.C., Ljung G.M.** Time series analysis: Forecasting and control. John Wiley & Sons; 2015: 712.

3. **Kaastra I., Boyd M.** Designing a neural network for forecasting financial and economic time series. Neurocomputing. 1996; (10):215–236.
4. **Winters P.R.** Forecasting sales by exponentially weighted moving averages. Management Science. 1960; 6(3):324–342.
5. **Engle R.F., Kroner K.F.** Multivariate simultaneous generalized ARCH. Econometric Theory. 1995; 11(1):122–150.
6. **Ryabko B., Astola J., Malyutov M.** Compression-based methods of statistical analysis and prediction of time series. Switzerland: Springer International Publishing; 2016: 144.
7. **Ryabko B.Ya.** Prediction of random sequences and universal coding. Problems of Information Transmission. 1988; 24(2):87–96.
8. **Chirikhin K.S., Ryabko B.Ya.** Experimental study of the accuracy of compression-based forecasting methods. Vestnik NSU. Series: Information Technologies. 2018; 16(3):145–158. (In Russ.)
9. **Tim S.** Prediction of infinite words with automata. Theory of Computing Systems. 2018; (62):653–681.
10. **Chirikhin K.S., Ryabko B.Ya.** Application of artificial intelligence and data compression methods to time series forecasting. Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Statistical Computation and Simulation — AMSA'2019". Novosibirsk. 2019: 553–560.