УДК 004.9 : 336.71 : 332.1

***© С.В. Коломийцева, В.А. Абрамсон, 2023***

**ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПОКАЗАТЕЛЕЙ СОСТОЯНИЯ БАНКОВСКОГО СЕКТОРА РЕГИОНОВ РФ[[1]](#footnote-1)**

*Коломийцева С.В.* – к.ф.-м.н, доцент кафедры "Высшая математика", e-mail: skolomy@gmail.com; *Абрамсон В.А*. – магистрант по направлению "Прикладная математика и информатика", (ЕНИ ДВГУПС), e-mail: vika9824@mail.ru

Методами интеллектуального анализа данных в статье выявляется взаимосвязь между экономическими и социальными показателями, строится прогноз значений экономических показателей, рассчитываются ошибки прогнозирования и выявляются наиболее подходящие методы прогнозирования для некоторых субъектов Российской Федерации.

Вata mining methods in the article reveal the relationship between economic and social indicators, build a forecast of the values of economic indicators, calculate forecasting errors and identify the most suitable forecasting methods for some subject of the Russian Federation.

*Ключевые слова:* интеллектуальный анализ данных, прогнозирование, временные ряды, экономические показатели, субъекты Российской Федерации.

**Введение.** Интеллектуальный анализ данных связан с поиском в данных скрытых нетривиальных закономерностей, которые позволяют получить новые знания об исследуемых данных. Интерес к методам анализа данных растет вместе с тем, как развиваются средства сбора и хранения данных, которые позволяют накапливать огромные объемы информации. Как итог, по всему миру перед специалистами из разных областей человеческой деятельности встает вопрос об обработке собираемых данных, а также превращения их в знания. В такой ситуации методы интеллектуального анализа данных приобретают особую актуальность. Их главная особенность заключается в обнаружении наличия и характера скрытых закономерностей в данных. Широкое и эффективное применение этих средств является одним из факторов успеха предприятия, особенно в условиях конкуренции.

В рамках проведения исследования нами был решен ряд задач по анализу региональных рынков финансовых услуг субъектов РФ на основании массива значений, отображающих объем рынка, а также долю банка относительно различных экономических показателей банковского сектора за период с 2017 по 2019г. Объектом исследования являлся универсальный коммерческий банк (КБ), предоставляющий розничные банковские продукты и услуги в рамках заданных территориальных границ. Среди экономических исследовался показатель оплаты услуг жилищно-коммунального хозяйства (ЖКХ), для которого строился прогноз значений до конца текущего финансового года средствами интеллектуального анализа.

**Подготовка данных к анализу**. Для подготовки данных к дальнейшей обработки может возникнуть необходимость в исправлении некоторой информации, для чего могут служить следующие причины: данные были введены некорректно; значения отсутствуют; данные содержат в себе пустые строки; значения находятся за пределами допустимых диапазонов; данные содержат значения, которые затрудняют их автоматизированный анализ.

Первые случаи можно решить либо удалением соответствующих строк, что может привести к снижению репрезентативности выборки, либо заменой значения параметра на среднее или граничное сверху или снизу, в зависимости от того, в каком диапазоне находится нехарактерное значение. Однако такая замена может снизить точность и адекватность построенной модели. Последний случай решается при помощи переразметки данных, что позволяет присвоить разным значениям одного объекта одно значение, тем самым, избавляясь от данных, затрудняющих автоматизированный анализ.

**Алгоритмы прогнозирования временных рядов**. В качестве инструмента для анализа показателей состояния банковского сектора регионов РФ нами используется надстройка интеллектуального анализа данных для Microsoft Office, в которой реализованы алгоритмы статистической обработки данных, основанные на анализе временных рядов, алгоритмы построения деревьев решений, алгоритм Байеса. Источником данных для анализа выступает электронная таблица Excel. Во время работы инструментов данные передаются на SQL Server, там обрабатываются, а результаты возвращаются и отображаются в Excel. Такое решение обусловлено широким распространением продукта MS Excel, для которого реализован большой круг дополнительных надстроек с бесплатной лицензионной политикой, а также достаточно прозрачной работой связки электронных таблиц MS Excel и SQL Server в области интеллектуального анализа данных.

Начиная с 2008 года в Analysis Services существует два алгоритма Microsoft Time Series для прогнозирования временных рядов: ARTXP (дерево авторегрессии с перекрестным прогнозированием) и ARIMA (авторегрессионное интегрированное скользящее среднее или, иначе, модель Бокса-Дженкинса).

Алгоритм ARTXP (дерево авторегрессии с перекрестным прогнозированием) был добавлен в Analysis Services в 2005 году. Он сочетает в себе метод авторегрессии с интеллектуальным анализом данных (дерево решений), поэтому уравнение прогноза может изменяться на основе определённых критериев. Этот алгоритм чаще применяется для краткосрочного прогнозирования временных рядов. Поддерживает перекрестное прогнозирование.

Деревья решений – это способ представления правил в иерархической и последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Под правилом здесь понимается логическая конструкция, которая представлена в виде «если …, то …». Этот метод применяют тогда, когда необходимо принимать последовательный ряд решений. Построение дерева решений начинается с более раннего решения, затем изображаются возможные действия и их последствия, затем снова принимается решение и далее до тех пор, пока все логические последствия результатов не будут исчерпаны.

Модель дерева авторегрессии (ART) представляет собой кусочно-линейную модель авторегрессии, которая может быть представлена в виде дерева. Листья дерева решений содержат линейные модели авторегрессии. На рисунке 1 показан пример модели ART, в которой есть три области, определенные с использованием переменной $Y\_{t+1}$, и каждый лист содержит модель AR, описываемую уравнением на листе.

Таким образом ветвление вдоль дерева работает в зависимости от прошлых значений для серии. Каждый лист представляет собой AR-модель для прогнозирования следующего значения временных рядов и описывается следующим соотношением:

$$f\left(y\_{t-p},…,t\_{t-1},θ\right)=N\left(m+\sum\_{j=1}^{p}b\_{j}y\_{t-j}, σ^{2}\right),$$

где $N\left(μ, σ^{2}\right)$ – нормальное распределение с параметрами $μ$ и $σ^{2}$, $y\_{t}$ – целевая переменная, для которой строится прогноз, $(y\_{t-p},…,t\_{t-1})$ – переменные предиктора, $θ=(m,b\_{1},..,b\_{p},σ^{2})$ – параметры модели [1].



*Рис. 1* Модель дерева авторегрессии.

Алгоритм ARIMA (интегрированное скользящее среднее с авторегрессией) был добавлен в Analysis Services в 2008 году для повышения точности и устойчивости долгосрочного прогнозирования и использует как авторегрессионные члены, так и члены скользящего среднего [2].

Алгоритм ARIMA [3]:

1. пусть *Y* обозначает исходных ряд;
2. пусть $y$ обозначает *d*-ю разность от *Y*:

если $d=1: y\_{t}=Y\_{t}-Y\_{t-1} $

если $d=2: y\_{t}=\left(Y\_{t}-Y\_{t-1}\right)-\left(Y\_{t-1}-Y\_{t-2}\right)=Y\_{t}- 2Y\_{t-1}+Y\_{t-2}$;

3) Общее уравнение имеет вид:

$$y\_{t}=μ+φ\_{1}y\_{t-1}+…+φ\_{p}y\_{t-p}-θ\_{1}ε\_{t-1}-…-θ\_{q}ε\_{t-q},$$

где $μ$ – константа, $φ\_{1}y\_{t-1},…φ\_{p}y\_{t-p}$ – авторегрессионная компонента порядка *p*; $θ\_{1}ε\_{t-1},…,θ\_{q}ε\_{t-q}$ – компонента скользящего среднего порядка *q*;

4) отменяем разность из пункта 2, чтобы получить прогноз:

если $d=1: Y\_{t}  =  y\_{t} + Y\_{t-1} $

если $d=2: Y\_{t}=y\_{t}+ 2Y\_{t-1} -Y\_{t-2}$.

По умолчанию алгоритм временных рядов Microsoft Time Series использует оба алгоритма и объединяет результаты (MIXED), что дает возможность управлять сочетанием алгоритмов, перенося акцент во временных рядах на краткосрочный или долгосрочный прогноз.

**Прогноз показателей и расчет ошибки прогнозирования.** Для решения задачи прогнозирования показателя ЖКХ были построены три модели соответственно имеющимся алгоритмам. Результаты построения моделей, отображающие исторические и прогнозируемые значения показателей ЖКХ Приморского, Хабаровского и Камчатского краев, представлены на рис. 2.



(*а*)



(*б*)



(*в*)

*Рис. 2* Результаты использования алгоритмов прогнозирования
по Приморскому краю (*а*), по Камчатскому краю (*б*), и по Хабаровскому краю (*в*).

В качестве ошибки прогнозирования была выбрана средняя абсолютная ошибка (MAE), которая рассчитывается как среднее значение абсолютных разностей между фактическими значениями и прогнозируемыми (таблица 1) [4]:

$$MAE=\frac{1}{N}\sum\_{i=1}^{N}\left|y\_{i}-\hat{y}\_{i}\right|, \left(1\right)$$

где $y\_{i}$ – фактическое значение ряда, $\hat{y}\_{i}$ – спрогнозированное значение ряда, N – количество временных меток.

Таблица 1 – Результаты расчетов ошибки прогноза

|  |  |  |  |
| --- | --- | --- | --- |
|  | ARTXP  | ARIMA  | MIXED  |
| Приморский край | 1,99 | 1,48 | 1,39 |
| Камчатский край | 13,27 | 15,26 | 14,38 |
| Хабаровский край | 2,90 | 3,65 | 2,64 |

**Заключение.** Использование смешанного алгоритма показывает, что величина его ошибок минимальна относительно остальных алгоритмов. Результаты вычислений прогноза, представленные в работе, хорошо согласуются с ожидаемыми значениями, однако несоответствие прогнозируемых значений фактическим данным для Камчатского края требует пристального внимания банковских специалистов для расследования точки изменения поведения модели.

**Благодарности.** Исследования проведены при поддержке РФФИ (гранты №16-07-00156, 16-37-00026 мол\_а) и ДВО РАН (гранты №15-I-4-071, №15-I-4-072).

**Библиографические ссылки**

1. Meek C., Chickering D. M., Heckerman D. Autoregressive Tree Models for Time-Series Analysis // Proceedings of the Second International SIAM Conference on Data Mining. 2002. 600 p.
2. Алгоритм временных рядов (Майкрософт) // Электронная документация по SQL Server. Microsoft Corporation. URL: <https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/microsoft-time-series-algorithm?view=sql-server-2017> (дата обращения: 10.07.2019).
3. Прикладные задачи анализа данных // Coursera. URL: <https://ru.coursera.org/learn/data-analysis-applications> (дата обращения: 10.07.2019).
4. Тихонов Э. Е. Методы прогнозирования в условиях рынка: учеб. пособие. Невинномысск, 2006. 221 с.
1. *VII Международная конференция «Информационные технологии и высокопроизводительные вычисления» (ITHPC-2023), Хабаровск, Россия, 11-13 сентября 2023 г. Сборник материалов трудов. Под ред. Смагин С.И.* [↑](#footnote-ref-1)