УДК 65.011.56

ПРИМЕНЕНИЕ СРЕДСТВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ (DATA MINING) ДЛЯ ИССЛЕДОВАНИЯ ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Ерисов В.Д.[1], Пекова Е.А.[1]

[1] ФГБОУ ВО «Московский политехнический университет», Москва, e-mail: mr.vovezzz@mail.ru

Интеллектуальный анализ данных (Data Mining) представляет собой процесс поиска практически полезных к использованию сведений в большом массиве данных. В интеллектуальном анализе данных применяются методы математического анализа для обнаружения закономерностей и взаимосвязей, существующих в данных, которые необходимы для построения гипотез или принятия решений. В основном такие закономерности и взаимосвязи довольно проблематично выявить при традиционном анализе данных, поскольку связи слишком сложны и запутанны из-за большого объема данных. В данной работе рассматривается практическое применение и исследование функциональных возможностей и особенностей инструментов Data Mining при анализе экономических показателей методом прогнозирования с помощью линейной регрессии используя аналитическую платформу Deductor Studio – программу, предназначенную для анализа информации из различных источников больших объемов данных и реализующую функции импорта, обработки, визуализации и экспорта данных. Кроме этого, предполагается исследовать зависимость чистой прибыли от основных средств, запасов, денежных средств и нераспределенной прибыли (непокрытого убытка), а также построить прогноз на следующий отчетный период. Показатель чистой прибыли будет рассматриваться как эндогенный фактор, поскольку получение прибыли обычно является главной целью всех видов предпринимательской деятельности.

Ключевые слова: Интеллектуальный анализ данных, Data Mining, Deductor Studio, регрессия, МНК.

APPLICATION OF MEANS OF INTELLECTUAL DATA ANALYSIS (DATA MINING) FOR THE INVESTIGATION OF ECONOMIC INDICATORS

Erisov V.D.[1], Pekova E.A.[1]

[1] Moscow Polytechnic University, Moscow, e-mail: mr.vovezzz@mail.ru

Intelligent data analysis (Data Mining) is the process of finding information that is practical to use in a large data array. In data mining, mathematical analysis methods are used to detect patterns and relationships that exist in the data, which are necessary for building hypotheses or making decisions. Such patterns and relationships are difficult to identify with traditional data analysis, since relationships are too complicated and confusing due to the large amount of data. This paper discusses the practical application and study of the functionality and features of using Data Mining tools when analyzing economic indicators using the linear regression prediction method using the analytical platform Deductor Studio, a program designed to analyze information from various data sources and implement the functions of import, processing, visualization and export data. In addition, it is supposed to investigate the dependence of net profit on fixed assets, stocks, cash and retained earnings (uncovered loss), as well as build a forecast for the next reporting period. The net profit indicator will be considered as an endogenous factor, since profit is usually the main goal of all types of business activities.

Keywords: Data Mining, Deductor Studio, Regression, OLS.

Процесс поиска и выявления закономерностей и зависимостей в практически полезных данных проводится методами интеллектуального анализа данных (Data Mining) путем исследования множества рассматриваемых объектов. В основной своей массе оно заключается в представлении исследуемых данных в табличном виде, где каждая строка соответствует одному из вариантов состояния рассматриваемого объекта, а столбцы содержат значения параметров, которые его характеризуют [1]. Выявление закономерностей и взаимосвязей предполагает наличие зависимой переменной. Она представляет из себя параметра, значение которого рассматривается как зависящее от других известных и заданных независимых переменных (параметров). Исходя из этого, задачей интеллектуального анализа данных (Data Mining) является определение взаимозависимости между исходными данными.

Рассмотрим основные задачи интеллектуального анализа данных (Data Mining) [2]:

1) Задача классификации. Она заключается в том, что для каждого варианта определяется категория или класс, которому он принадлежит.

2) Задача регрессии. Оно во многом схожа с задачей классификации, но в ходе ее решения производится поиск шаблонов для определения числового значения.

3) Задача кластеризации (сегментации). Она заключается в делении множества объектов на группы (кластеры) схожих по параметрам.

4) Задача определения взаимосвязей (поиска ассоциативных правил). Она также называемая задачей поиска ассоциативных правил, заключается в определении часто встречающихся наборов объектов среди множества подобных наборов.

5) Анализ последовательностей или сиквенциальный анализ. Целью данного анализа является обнаружение закономерностей в последовательностях событий.

6) Анализ отклонений. Он позволяет отыскать среди множества событий те, которые существенно отличаются от нормы.

Современные принципы организации и управления экономикой требуют от специалиста знания новых методов анализа и прогнозирования реальных экономических процессов, основанных на интеллектуальном анализе данных. Одним из инструментов интеллектуального анализа данных экономических показателей является аналитическая платформа Deductor Studio — программа, позволяющая проводить всесторонний анализ финансово-хозяйственной деятельности предприятия и прогнозировать его экономические показатели.

Архитектура аналитической платформы Deductor Studio позволяет реализовать все этапы построения аналитической системы: от создания хранилища данных предприятия до настройки корпоративной аналитической отчетности [3]. Deductor Studio позволяет аналитику автоматизировать рутинные операции по обработке данных и сосредоточиться на

интеллектуальной работе: формализация логики принятия решений, построение моделей, прогнозирование.

В результате работы данного компонента аналитической платформы Deductor Studio строится линейная модель данных. Применяется следующий алгоритм построения модели.

Пусть имеется набор входных значений Xi, где i = 1 ... n, т.е X = {x1,x2, ... , xn}. Тогда можно указать такой набор выходных значений Yj (j = 1 … m), который будет соответствовать линейной комбинации входных значений с коэффициентами ai (i = 1 … n):

$$[1, x1, x2, \dots , xn] [a0, a1, a2, \dots, an] = [y1, y2 ,\dots, ym] \quad (1.1)$$

Если для простоты предположить, что выходное значение одно, то можно записать:

$$a0 + x1a1 + x2a2 + \dots + xnan = y \quad (1.2)$$

Таким образом, задача сводится к подбору коэффициентов ai. Их оценка производится путем метода наименьших квадратов (МНК).

Однако, следует отметить, что использования метода линейной регрессии предназначена для поиска только линейных зависимостей в данных. В случае, когда зависимость нелинейная, то построенная модель, скорей всего, будет недостаточно точной. Это можно увидеть на диаграмме рассеяния, так как разброс прогнозных значений величины будет сильно велик относительно действительных значений. В данном случае необходимо использовать более эффективные алгоритмы, например, нейронные сети.

В первую очередь, для того, чтобы начать использовать нейросеть, ее необходимо обучить. Задача обучения здесь равносильна задаче аппроксимации функции, то есть восстановление функции по отдельно взятым ее точкам – таблично заданной функции [4]. В силу этого, для обучения нейросети необходимо подготовить обучающую выборку, то есть таблицу с входными значениями и соответствующими им выходными значениями. Такая таблица поможет нейросети самой обнаружить взаимосвязи и зависимости выходных полей от входных. Далее, подавая на вход нейросети некоторые значения, можно использовать эти взаимосвязи и зависимости. На выходе будут восстановлены зависимые от них значения, причем на вход можно подавать значения, на которых нейросеть не обучалась [5].

В работе будет исследоваться зависимость чистой прибыли от основных средств, запасов, денежных средств и нераспределенной прибыли (непокрытого убытка), а также построится прогноз на следующий отчетный период.

Под самой чистой прибылью понимается часть балансовой прибыли предприятия, остающаяся в его распоряжении после уплаты налогов, сборов, отчислений и других обязательных платежей в бюджет. Чистая прибыль используется для увеличения оборотных средств предприятия, формирования фондов и резервов, и реинвестиций в производство.

Данный показатель будет рассматриваться как эндогенный фактор, поскольку получение прибыли обычно является главной целью всех видов предпринимательства.

Чистая прибыль Y в нашей модели будет зависит от следующих показателей:

1. Основные средства – это материальные активы, которые предприятие содержит с целью использования их в процессе производства или поставки товаров, предоставления услуг, сдачи в аренду другим лицам или для осуществления административных и социально-культурных функций.

Зависимость между основными средствами и чистой прибылью компании показывает эффективность использования имеющихся ресурсов, которые в свою очередь напрямую влияют на чистую прибыль организации, поэтому данный фактор включаем в модель как экзогенный фактор X1.

2. Запасы – это активы, используемые в качестве сырья, материалов и т. п. при производстве продукции, предназначенной для продажи (выполнения работ, оказания услуг), приобретаемые непосредственно для перепродажи, а также используемые для управленческих нужд организации.

Управление запасами для получения прибыли является одной из наиболее сложных и главных задач руководства организации. Запасы зачастую являются самыми крупными оборотным активом предприятия. Поэтому сумма, отнесенная на запасы, напрямую влияет на сумму чистой прибыли, значит данный фактор включаем в модель как экзогенный фактор X2.

3. Денежные средства - это наличные деньги в кассе организации, средства на банковских счетах и средства, воплощенные в денежных документах.

Денежные средства участвуют в процессе производства и управления, а значит увеличение или уменьшение чистой прибыли пропорционально их росту или уменьшению, поэтому данный фактор включаем в модель как экзогенный фактор X3.

4. Нераспределенная прибыли (непокрытый убыток) - это прибыль (убыток) компании, фирмы, акционерного общества, остающаяся после уплаты налогов и выплаты дивидендов, используемая для реинвестирования, на нужды развития или же убытки, который понесла фирма за время ведения всех видов деятельности за предыдущие отчетные периоды

Нераспределенная прибыли и непокрытый убыток влияют на чистую прибыль, потому что, например, прибыль, не инвестированная в активы фирмы, может сказаться на прибыли следующих периодов. Точно так же из-за непокрытого убытка средства придется направлять на его погашение, что скажется на значении чистой прибыли следующих периодов. Поэтому данный фактор включаем в модель как экзогенный фактор X4.

Для проведения интеллектуального анализа выбранных экономических показателей необходимо импортировать текстовый файл с их значениями за разные периоды непосредственно в аналитическую платформу Deductor Studio (Рисунок 1).

Таблица

1 / 22

№	y	x1	x2	x3	x4
1	-2851	102	390	8342	-4407
2	-2647	109	404	8393	-4350
3	57	2045	375	2661	-4031
4	63	2051	397	2688	-4012
5	201	747	341	4650	-3830
6	211	752	343	4657	-3803
7	203	746	340	4645	-3832
8	209	753	344	4662	-3801
9	4007	992	24031	1874	384
10	4208	997	24112	1918	393
11	69	2440	26200	1940	1258
12	63	2481	26907	1978	1344
13	-1867	3812	29238	3386	-675
14	-2073	3842	29352	3401	-664
15	-1913	3819	29243	3382	-672
16	-2037	3835	29347	3405	-667
17	865	3204	22092	5795	120
18	719	3220	22121	5817	126
19	892	3211	22098	5793	122
20	692	3213	22114	5819	124
21	-10168	1526	22008	4873	-10711
22	-11801	1651	22023	4994	-11013

Рисунок 1 - Исходная таблица данных в программе Deductor Studio

Далее на основании импортированных данных необходимо провести анализ методом прогнозирования с помощью линейной регрессии. Для того, чтобы его провести, нужно в меню найти раздел «Мастер обработки» и выбрать в качестве обработки данных «Линейную регрессию» (Рисунок 2).

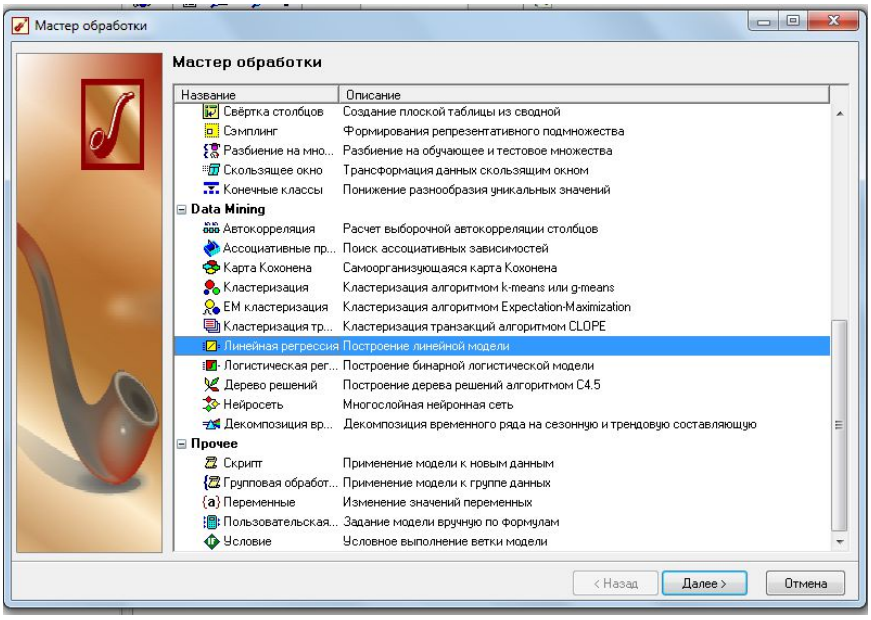


Рисунок 2 - Мастер обработки сценариев

После выбора нужного метода обработки данных необходимо задать назначение для исходных столбцов (Рисунок 3).

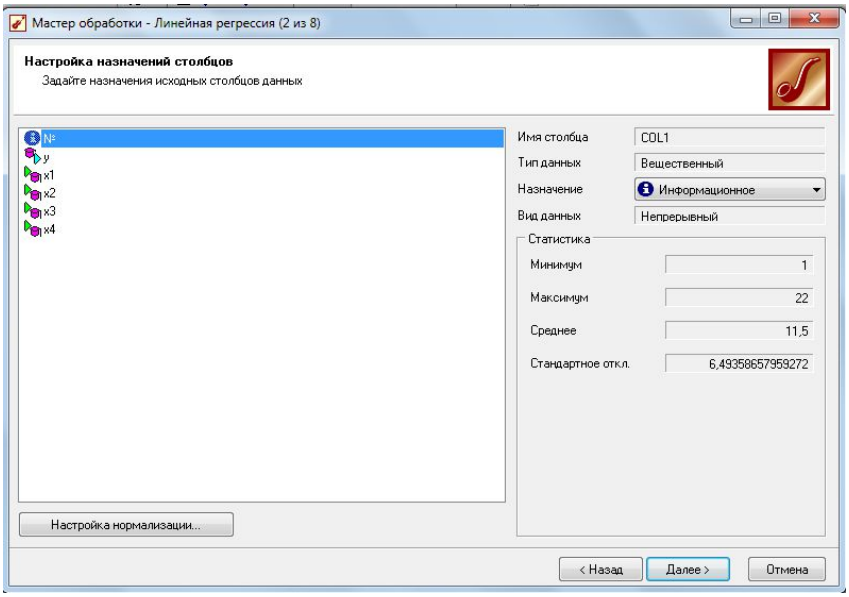


Рисунок 3 - Настройка назначения столбцов

В данном случае для номеров наблюдения было выбрано назначение «Информационное», так как они являются вспомогательными, то есть их не следует использовать при обработке. Для всех значений X укажем назначение «Входное», так как предполагается, что они влияют на Y. Соответственно, для значения Y присвоим назначение «Выходное»

На следующем шаге происходит настройка обучающего и тестового множеств и способ разложения исходного множества данных (Рисунок 4).

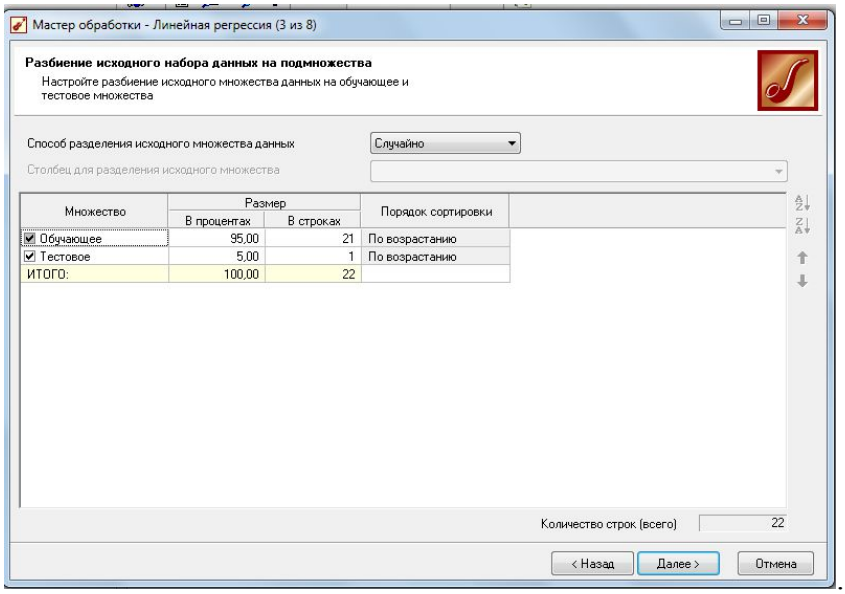


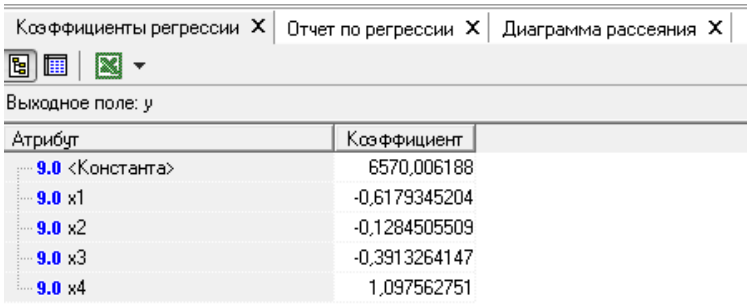
Рисунок 4 - Разбиение исходного множества на подмножества

Далее установщик позволяет осуществить ограничение диапазона входных значений. Данный шаг оставим без изменений.

После выполнения процесса выберем в качестве способа отображения «Коэффициенты регрессии», «Отчет по регрессии» и «Диаграмму рассеяния».

По рассчитанным коэффициентам регрессии (Рисунок 5) можно сделать следующие выводы:

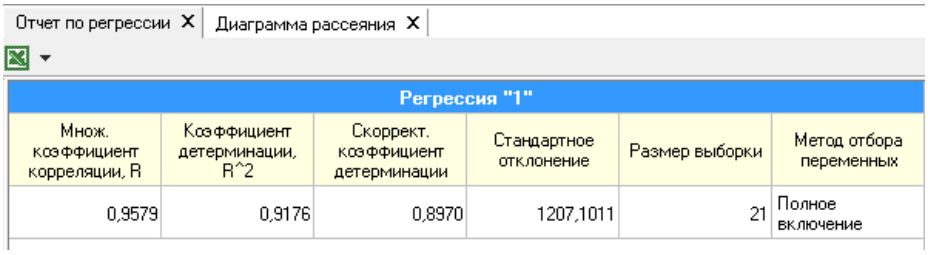
1. Чистая прибыль предприятия без учета включенных в модель факторов будет составлять 6570,1 тыс. руб.
2. При увеличении основных средств на 1 тыс. руб. при неизменном уровне других факторов чистая прибыль уменьшится на 0,61 тыс. руб.
3. При увеличении запасов на 1 тыс. руб. при неизменном уровне других факторов чистая прибыль уменьшиться на 0,13 тыс. руб.
4. При увеличении денежных средств на 1 тыс. руб. при неизменном уровне других факторов чистая прибыль уменьшится на 0,39 тыс. руб.
5. При увеличении суммы нераспределенной прибыли на 1 тыс. руб. при неизменном уровне других факторов чистая прибыль увеличится на 1,1 тыс. руб.



Атрибут	Коэффициент
9.0 <Константа>	6570,006188
9.0 x1	-0,6179345204
9.0 x2	-0,1284505509
9.0 x3	-0,3913264147
9.0 x4	1,097562751

Рисунок 5 - Коэффициенты регрессии

Благодаря анализу регрессии (Рисунок 6) можно сделать выводы относительно выбранных нами параметров, а именно: значение коэффициента детерминации R-квадрат, который равен 0,918, говорит нам о том, что у чистой прибыли сильная связь с выбранными факторами. Следовательно, в 91,8 % случаев изменение чистой прибыли компании связано с изменением включенных в модель факторов.



Регрессия "1"					
Множ. коэффициент корреляции, R	Коэффициент детерминации, R^2	Скоррект. коэффициент детерминации	Стандартное отклонение	Размер выборки	Метод отбора переменных
0,9579	0,9176	0,8970	1207,1011	21	Полное включение

Рисунок 6 - Отчет по регрессии

Кроме этого, из диаграммы рассеянья (Рисунок 7) видно, что разброс между эталонными значениями выходного поля и значениями, рассчитанными моделью, достаточно невелик. Из этого можно сделать вывод, что временной ряд хорошо укладывается в линейную

модель и, следовательно, на основании этой модели можно строить прогноз на будущие периоды времени.

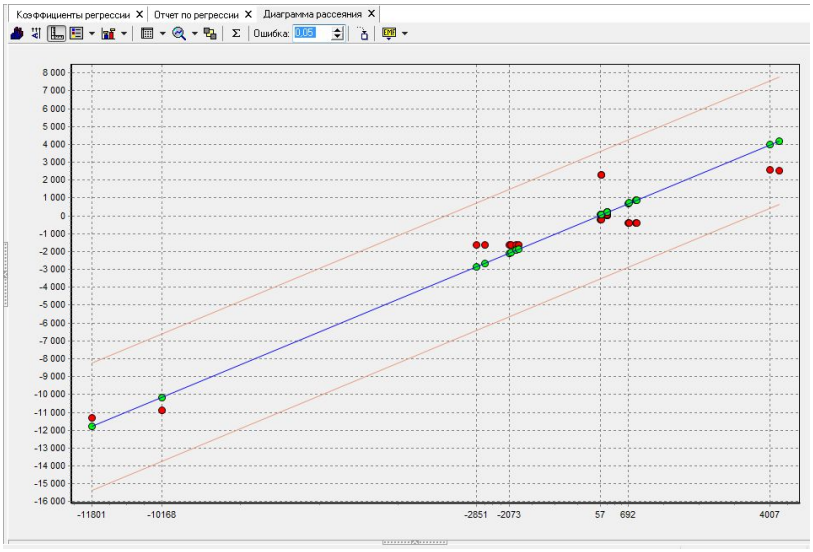


Рисунок 7 - Диаграмма рассеивания

Следующим и последним шагом является прогнозирование. Для построения прогноза чистой прибыли на следующий период необходимо на основании построенной регрессии в разделе «Мастере обработки» выбрать в качестве обработки данных «Прогнозирование». В результате чего в исходную таблицу добавится новая строка с прогнозным значением чистой прибыли на следующий период (Рисунок 8).

Таблица

23 / 23

№	y	x1	x2	x3	x4	Шаг прогноза
1	-2851	102	390	8342	-4407	
2	-2647	109	404	8393	-4350	
3	57	2045	375	2661	-4031	
4	63	2051	397	2688	-4012	
5	201	747	341	4650	-3830	
6	211	752	343	4657	-3803	
7	203	746	340	4645	-3832	
8	209	753	344	4662	-3801	
9	4007	992	24031	1874	384	
10	4208	997	24112	1918	393	
11	69	2440	26200	1940	1258	
12	63	2481	26907	1978	1344	
13	-1867	3812	29238	3386	-675	
14	-2073	3842	29352	3401	-664	
15	-1913	3819	29243	3382	-672	
16	-2037	3835	29347	3405	-667	
17	865	3204	22092	5795	120	
18	719	3220	22121	5817	126	
19	892	3211	22098	5793	122	
20	692	3213	22114	5819	124	
21	-10168	1526	22008	4873	-10711	
22	-11801	1651	22023	4994	-11013	
22	-11320,8128742005	1651	22023	4994	-11013	1

Рисунок 8 – Прогнозное значение

Так как значения нераспределенной прибыли в таблице за последнее время отрицательное, то мы имеем дело с непокрытым убытком. По уравнению регрессии увеличение этого фактора увеличивает прибыль, но по последним данным вместо прибыли организация несет убыток, а значит увеличение значения непокрытого убытка увеличит убыток в следующем году. Отсюда следует, что надо уменьшить значение непокрытого

убытка, чтобы организация начала получать прибыль или сократила численное значение убытка. Это она может сделать за счет резервного фонда, добавочного капитала или привлечения средств из вне, то есть инвестиций или целевых взносов учредителей. Кроме того, увеличение расходов на материальные активы тоже уменьшает прибыль. Исходя из этого, для увеличения чистой прибыли или сокращения убытка необходимо сократить значения этих факторов.

На основании этого прогноза можно сделать вывод, что убыток в следующий квартал сократиться с 11801 тыс. руб. до 11321 тыс. руб. Однако в связи с тем, что для получения прогноза на большое число шагов используются не реальные, а вычисленные с использованием модели данные, ошибка такого прогноза может быть довольно значительной. Поэтому при построении прогноза на длительный период величина погрешности возрастает с каждым новым шагом, что существенно снижает ценность прогноза. Поэтому для более точных прогнозов и более серьезного анализа данная модель нуждается в дополнительных данных.

Список литературы

1. Айвазян С.А. Методы эконометрики: учебник. – М.: Магистр: ИНФРА-М, 2014.
2. Богомолова, М.А. Интеллектуализация систем принятия управленческих решений по взаимодействию с клиентами / М.А. Богомолова // Экономика и социум. –2014. – No3-1 (12). –С. 264-266.
3. Васильев В.П. Информационно-аналитические системы. Практикум на ПК.МФ МЭСИ -2010.
4. Кремер Н.Ш., Путко Б.А. Эконометрика. Учебник. М.: ЮНИТИ, 2013.
5. Курзаева Л.В. Введение в методы и средства получения и обработки информации для задач управления социальными и экономическими системами: учеб. пособие/Л.В. Курзаева, И.Г. Овчинникова, Г.Н. Чусавитина. -Магнитогорск: Магнитогорск. гос. техн. ун-та им. Г.И. Носова, 2016. -118 с.