Анализ и прогнозирование для рынков труда на основе онлайн-данных

В.С. Гиоргашвили, М.А. Бакаев

***Аннотация* — Проблема неполноты данных весьма характерна при проведении социологических, экономических и статистических исследований с использованием онлайн-данных. Возможными причинами неполноты данных могут быть: ошибки и изменения на площадках-источниках данных, сбои и ошибки в работе инструментов, осуществляющих сбор данных и т.п. Поскольку для прогнозирования ситуации на рынке труда наличие пропусков в данных обычно нежелательно, то предпочтительное решение – заполнить недостающие значения с применением подходящего метода, не приводящего к искажению результатов. В данной статье представлен обзор методов устранения неполноты данных и описано применение метода k-средних для заполнения пропусков в собранных нами онлайн-данных. Для сбора данных по рынкам труда ряда регионов Сибирского ФО использовалась специализированная веб-майнинговая программная система, которая с 2011 г. и по настоящий момент извлекла и сохранила для анализа более 2 млн. уникальных наборов данных (вакансий и резюме). Эффективность использованного метода проверяется посредством сравнения полученных результатов (восстановленных средних заработных плат и количества вакансий) с данными, которые были позднее собраны с использованием механизмов API онлайн-площадок. Также мы применяем интегрированную модель авторегрессии скользящего среднего (ARIMA) для прогнозирования спроса на рынке труда относительно ИТ-специалистов. Сравнение предсказанных и дополнительно собранных в 2018 г. данных позволяет сделать вывод о применимости предложенной модели для мониторинга и управления на рынке труда.**

***Ключевые слова*— качество данных, пропуски в данных, веб-майнинг, рынок труда, k-среднее, прогнозирование рынка труда**

I. Введение

В настоящее время активно разрабатываются и применяются программные системы, предназначенные для сбора и обработки онлайн-данных [1]. Среди систем с открытым исходным кодом можно, например, отметить платформы Nutch and Solr, основанные на проекте Apache Lucene. Nutch извлекает, анализирует и индексирует для каждого URL-адреса все связанные ресурсы в соответствии с рядом ограничений, таких как глубина ссылки с корневой страницы, которая должна сканироваться, и максимальное количество страниц, которые будут извлекаться. Solr используется для поиска любых типов данных и для поиска веб-страниц. К основным функциям этой платформы относятся полнотекстовый поиск, фасетный поиск, динамическая кластеризация, интеграция с базами данных и расширенная обработка документов. Широко также применяются специализированные (как правило, коммерческие) программные инструменты для анализа настроений, сбора отзывов из социальных сетей и т.д.

В ходе использования таких систем при проведении разнообразных социологических, экономических и статистических исследований типична ситуация, когда необходимо обработать пропуски в массивах данных. Так, при работе с данными по рынку труда, собранными из онлайн-источников, нами была отмечена неполнота данных о значении средней заработной платы для отдельных сфер деятельности и регионов. То есть система сбора данных в отдельные периоды не собирала информацию с онлайн-площадок и, следовательно, база данных системы не пополнялась. Это, как правило, связано с изменением структуры программного кода веб-страниц площадок или, в некоторых случаях, со сбоем в работе самой системы.

Используя информацию, представленную в семантически организованном виде, т.е. собранную с помощью API источников, неизбежны некоторые недостатки данного способа сбора данных. К ним относится ограничение на количество одновременных запросов и на количество обращений, которые приложение может делать в единицу времени. Кроме того, необходимо постоянно отслеживать изменения в API и обновлять приложение по сбору данных, причем некоторые площадки предоставляют важные данные не в полном объеме или на платной основе.

Сбор данных с веб-страниц, который изначально был реализован в разработанной нами программной системе, вполне оправдывает себя на практике, но имеет собственные недостатки. Так, при обработке данных система исключает все ненужные поля веб-страниц (шапка и подвал сайта, новостные и рекламные баннеры и пр.), оставляя только контент. Проблемы извлечения информации в данном случае могут быть связанны с изменением структуры веб-станиц в контентной части сайта. В БД системы каждому полю веб-страницы соответствует свой маркер (обычно идентификатор тега div), но изменение структуры сайта может приводить к тому, что поля перестают соответствовать заголовкам, что может приводить к сбою в сборе данных. Во избежание этого необходимо периодически отслеживать корректность осуществления сбора (проверять БД

Статья получена 22.10.2018.
Работа выполнена при финансовой поддержке РФФИ/РГНФ в рамках научного проекта № 17-32-01087 ОГН а2.
Гиоргашвили Виктория Сергеевна, Новосибирский государственный технический университет, магистр, (e-mail: vikagiorgashvili@mail.ru).
Бакаев Максим Александрович, Новосибирский государственный технический университет, доцент, канд. техн. наук, с.н.с. (e-mail: maxis81@gmail.com).

вручную или реализовать скрипт мониторинга с соответствующей логикой) и производить реконфигурации полей, либо вносить элементы искусственного интеллекта, чтобы система автоматически подстраивалась под относительно мелкие изменения веб-страниц.

Разработанная нами программная система (работает с 2011 г., изначально создавалась для нужд Комитета по труду мэрии г. Новосибирска) предназначена для поддержки принятия решений при управлении рынком труда. Система может автоматически собирать данные из указанных источников в интернете, которые открыто и массово публикуют вакансии, связанные с объявлениями о работе. Архитектура системы включает в себя три основных уровня:

1. Модуль сбора данных, отвечающий за доступ к исходным веб-сайтам и извлечение данных с веб-страниц.

2. Модуль обработки, отвечающий за структурирование данных. В настоящее время он извлекает информацию, связанную с вакансиями и резюме, и их конкретные свойства, а также классифицирует рабочие места по отраслям.

3. Модуль анализа, напрямую отвечающий за поддержку принятия решений и предоставление возможности для составления отчетов, фильтрации, уведомлений и т.д. [2, с. 26].

Рассматриваемая система для сбора, структурирования и анализа онлайн-данных (применяемая для мониторинга рынка труда в нескольких сибирских регионах, см. [3, с. 18]) не имеет встроенных механизмов для коррекции пропусков в данных, в связи с чем возникла необходимость подбора метода для устранения неполноты данных. В данной статье мы приводим краткий обзор методов, существующих в этой области, применяем метод k-средних для заполнения «пробелов» в данных по рынку труда и оцениваем качество полученных результатов, сравнивая их с дополнительно собранными системой данными.

II. Обзор методов устранения неполноты данных

На сегодняшний день существует множество методов, позволяющих устранить неполноту данных, каждый из которых имеет свои преимущества и недостатки:

1. Исключение из таблицы строк с пропусками. Метод применяется в случае с таблицей большой размерности и при незначительном количестве пропусков. В противном случае такой метод приводит к смещению оценок, потому как строки с пропущенными значениями содержат информацию, необходимую для анализа. Главным недостатком данного метода является потеря информации при изъятии неполных данных.

2. Заполнение пропусков средними по столбцу значениями. Применение данного метода целесообразно только в том случае, когда пропуски в данных по переменным случайны и сам механизм пропусков несущественен. Недостатками такого метода являются вносимые изменения в распределение данных и уменьшение дисперсии.

3. Метод ближайших соседей. Суть метода состоит в поиске строк таблицы, которые являются ближайшими по определенному критерию к строке с пропуском. Для его заполнения значения переменной (в установленном столбце) в соседних строках усредняются с конкретными весовыми коэффициентами, которые обратно пропорциональны расстоянию к строке, в которой есть пропуск. Такой метод точнее предыдущего, но он практически неприменим в случае большого количества пропусков, т.к. опирается на существование связей между строками в таблице.

4. Метод регрессии. По имеющимся данным осуществляется построение уравнения множественной линейной регрессии и вычисляются пропущенные значения переменных. Метод нельзя применить в случае, когда количество пропусков в строке больше одного, поскольку это приводит к множеству решений, и вместе с тем его точность является невысокой, поскольку в реальных задачах зависимости могут быть нелинейными [4, с. 52].

5. Алгоритм ZET. Основная идея алгоритма состоит в подборе «компетентной матрицы». Оперируя данными из этой матрицы, находят параметры зависимости, применяемой для прогнозирования пропущенного значения. Недостаток алгоритма заключается в его локальности, потому как для вычисления пропущенного значения используется лишь некоторая часть данных таблицы, а не все. По данным компонентной матрицы строится функциональная зависимость прогнозируемого значения от соответствующего значения в компетентной матрице, на основе которой затем прогнозируется значение пропуска [5, с. 264].

6. Resampling метод. Данный метод является итерационным и имеет две модификации, основанные на построении регрессионных моделей с последующим усреднением полученных оценок для пропущенных значений. Преимущество такого метода – повторное использование исходных данных, т.к. увеличение числа подвыборок позволяет наиболее точно использовать исходную информацию. Недостаток данного метода заключается в том, что объем новой информации уменьшается для каждой новой подвыборки, поскольку увеличивается вероятность того, что данные элементы выборки уже были выбраны раньше [4, с. 56].

7. Метод k-средних. При использовании данного метода, так же как в случае ближайших соседей, предполагается, что близкие по одним признакам строки должны быть близки и по другим признакам. Однако отличие метода состоит в том, что здесь осуществляется поиск не ближайших соседей для каждой строки с пропущенными значениями, а используется информация о центре кластера, куда попала конкретная строка с пропуском. Для разбиения на кластеры необходима начальная инициализация пропущенных значений [6, с. 282]. При использовании этого метода выполняется инициализация пропущенных значений с помощью замены средним значением по признаку, кластеризация производится методом k-средних. Пропущенные значения заменяются на соответствующие им значения центра кластера, в который попала каждая строка с пропуском. Этот алгоритм выполняется в течение нескольких итераций до сходимости или по достижению максимального

заданного числа итераций.

Для устранения неполноты имеющихся данных по рынку труда наиболее подходящим методом был признан метод k-средних, поскольку основным достоинством данного алгоритма является высокая скорость выполнения и эффективность в сравнении с другими методами, особенно в случае, когда речь идет о работе с крупными наборами данных.

III. Применение метода k-средних для заполнения пропусков в онлайн-данных по рынку труда

Основной причиной неполноты данных при применении системы онлайн-мониторинга рынка труда (подробное описание системы см. в [3, с. 17]) являлась изменчивость источников данных – площадок, служащих для размещения объявлений о вакансиях и резюме. При использовании метода k-средних объекты объединяются в кластеры так, что в один кластер попадут максимально схожие объекты, а объекты различных классов будут максимально отличаться друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, характеризующих объекты.

Неполные данные о значениях средней заработной платы по вакансиям и резюме и их количестве рассчитывались на основе имеющихся данных (использовался пакет Statistica). Для расчета были взяты данные по Новосибирской области за 2 полугодие 2016 года, представленные в таблице 1 (сбор за 2016 год). Для этого региона отсутствовали значения для таких сфер деятельности как «Страхование», «Торговля оптовая», «Недвижимость», «Телекоммуникация и связь», «Работа для студентов», «ТЭК, энергетика, добыча сырья», «Работа дома», «Туризм, гостиничное дело» и «Сельское хозяйство».

Таб. 1. Данные о вакансиях по Новосибирской области за 2 полугодие 2016 года

Сфера деятельности	Сбор за 2016 год		Сбор за 2017 год (новый)	
	Вакансий в среднем за неделю	Средняя зарплата, руб./мес.	Вакансий в среднем за неделю	Средняя зарплата, руб./мес.
Страхование	0	0	0,08	19250
Спорт, красота, здоровье	1,04	29806	1,22	28833
Рабочие профессии	0,96	25449	1,85	29201
Прочее	9,63	34894	11,81	29846
Государственная служба	0,27	29000	0,50	35955
Торговля розничная	6,15	35253	6,96	32432
Торговля оптовая	0	0	0,26	37000
Рестораны, кафе, общепит	4,85	23278	4,85	23278
Транспорт	4,41	34541	5,93	32077
Промышленность непищевая	0,04	40000	0,04	40000
Высший менеджмент	0,08	60000	0,19	63375
Логистика, склад, закупки	7,44	28391	8,22	27743
Строительство, архитектура	3,67	34152	4,63	35707
Бухгалтерия, финансы, банки	3,81	31801	5,65	30654
ИТ и Интернет	10,56	38343	11,67	37399
Маркетинг, реклама, PR	2,00	31908	2,41	26594
Сфера услуг	1,38	31625	1,96	25844
Охрана и безопасность	2,22	20533	2,89	20248
Медицина и формация	1,93	27569	2,96	24542
Персонал офиса, АХО	4,15	24531	4,50	24029
Недвижимость	0	0	0	0
Юриспруденция	0,81	24577	0,96	25390
Образование, наука, языки	2,74	19212	3,63	19486
Продажа услуг	11,52	30805	12,19	31007
Работа для студентов	0	0	0,19	28750
Персонал для дома	1,04	27386	1,19	26060
Промышленность пищевая	13,41	32112	14,74	31656
Телекоммуникация и связь	0	0	0,15	24667
Полиграфия, издательства, СМИ	0,15	30000	0,42	33125
ТЭК, энергетика, добыча сырья	0	0	0,04	40000
Работа дома	0	0	0	0
Туризм, гостиничное дело	0,19	0	0,38	18600
Временная работа	0,12	34250	0,12	34250
Кадровые службы, HR	0,85	29850	1,00	27547
Сельское хозяйство	0	0	0,08	27500
Дизайн, творческие профессии	1,50	25170	1,79	26353
Всего за неделю/средняя ЗП:	96,92	31299	115,46	30128

Сначала восстанавливались данные по вакансиям. С помощью функции «Иерархическая классификация» было определено количество кластеров. Объектами в данном случае были выбраны наблюдения (строки) – сферы деятельности. В качестве меры близости использовалось Евклидово расстояние – геометрическое расстояние между переменными в многомерном пространстве, которое вычисляется по исходным, а не по стандартизованным данным. Это обычный способ его вычисления, обладающий определенными преимуществами (например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться выбросом).

Исходя из визуального представления результатов, было выявлено, что сферы образуют 3 естественных кластера. Согласно методу k-средних, вычисления

начинались с k случайно выбранных наблюдений (k=3), которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами. После изменения состава кластера вычисляется новый центр тяжести, чаще всего как вектор средних значений по каждому параметру. Алгоритм продолжается до тех пор, пока состав кластеров не перестанет меняться [7, с. 185]. В результате данные по вакансиям разбились по кластерам (Рис. 1-3).

Элементы кластера номер 1 и расстояния до центра кластера. Кластер содержит 9 набл.

	объедин.
Страхование	5616,80
Торговля оптовая	4318,87
Высший менеджмент	6546,02
Строительство, архитектура	2593,74
Бухгалтерия, финансы, банки	4319,47
ИТ и Интернет	7384,76
Недвижимость	4786,33
Продажа услуг	3935,21
Телекоммуникация и связь	4782,68

Рис. 1. Первый кластер, содержит 9 наблюдений

Элементы кластера номер 2 и расстояния до центра кластера. Кластер содержит 15 набл.

	объедин.
Спорт, красота, здоровье	510,46
Рабочие профессии	358,56
Прочее	1793,92
Транспорт, автобизнес	275,24
Промышленность непищевая	1612,12
Логистика, склад, закупки	839,35
Маркетинг, реклама, PR	2591,31
Медицина и формация	2357,47
Юриспруденция	207,90
Промышленность пищевая	1024,56
Полиграфия, издательства, СМИ	1837,97
ТЭК, энергетика, добыча сырья	3433,97
Кадровые службы,HR	1377,02
Сельское хозяйство	566,47
Дизайн, творческие профессии	949,09

Рис. 2. Второй кластер, содержит 15 наблюдений

Элементы кластера номер 3 и расстояния до центра кластера. Кластер содержит 11 набл.

	объедин.
Государственная служба	179,40
Торговля розничная	1701,05
Рестораны, кафе, общепит	1291,97
Сфера услуг	181,08
Охрана и безопасность	1168,53
Образование, наука, языки	1046,05
Работа для студентов	144,67
Персонал для дома	37,62
Работа дома	1643,34
Туризм, гостиничное дело	1521,55
Временная работа	2931,74

Рис. 3. Третий кластер, содержит 11 наблюдений

Для того чтобы заполнить отсутствующие значения в соответствующих строках, было использовано среднее значение по каждому кластеру, в которые попали сферы с отсутствующими значениями параметров. Таким образом, можно предположить, что в среднем за неделю публиковалось 3,3 вакансий в сферах «Страхование», «Торговля оптовая» и «Недвижимость», а средняя заработная плата по этим сферам составила 21678 руб./месяц (рис. 4).

Описат.статистики для кластера 1 Кластер содержит 9 набл.

перемен.	Среднее	Стандарт отклон.	Дисперс.
Вакансии в среднем за неделю	3,29	3,64	19,40
Сдедняя з/п, руб/мес	21677,91	19269,25	440824581,00

Рис.4. Среднее количество вакансий и средняя заработная плата для первого кластера

Аналогичным образом заполнялись отсутствующие данные для остальных сфер деятельности, которые попали во 2 и 3 кластеры. Таким образом, в ходе вычислений были получены следующие результаты:

1. В среднем за неделю публиковалось 5,7 вакансий в сферах «ТЭК, энергетика, добыча сырья» и «Сельское хозяйство», а средняя заработная плата составила 25170 руб./месяц.
2. В сферах «Работа для студентов», «Работа дома», «Телекоммуникация и связь» количество опубликованных в среднем за неделю вакансий составило 1,7, а значение средней заработной платы – 20049 руб./месяц
3. В сфере «Туризм, гостиничное дело» отсутствовало значение средней заработной платы, в ходе вычислений этот показатель составил 21678 руб./месяц.

В сентябре 2017 года в работу системы были внесены некоторые изменения для запуска сбора данных на основе API онлайн-площадок. Таким образом, в конце 2017 года удалось восстановить некоторые пропуски в данных за предыдущие периоды. Но вместе с тем, как можно заметить из таблицы 1 (сбор за 2017 год), произошли изменения и в некоторых значениях по остальным сферам деятельности. Если сравнивать значения показателей, рассчитанные методом k-средних и данные собранные системой новым способом, то можно заметить что есть расхождения в значениях (таблица 2).

Таб. 2. Сравнение значений показателей, полученных разными способами

Сфера деятельности	**Данные, собранные на конец 2017 года**		**Данные собранные методом k-средних**		**Отклонение**	
	Вакансий в среднем за неделю	**Средняя зарплата, руб./мес.**	**Вакансий в среднем за неделю**	**Средняя зарплата, руб./мес.**	**Вакансии**	**Зарплата**
Страхование	0,08	19250	3,3	21678	-97,6%	-11,2%
Торг.оптовая	0,26	37000	3,3	21678	-92,1%	+70,7%
Недвижимость	0	0	3,3	21678		
Работа для студентов	0,19	28750	1,7	20049	-88,8%	+43,4%
Телекоммуникация и связь	0,15	24667	3,3	21678	-95,5%	+13,8%
ТЭК, энергетика, добыча сырья	0,04	40000	5,7	25170	-99,3%	+58,9%
Работа дома	0	0	1,7	20049		
Туризм, гостиничное дело	0,38	18600	0,19	20049	+100%	-7,2%
Сельское хозяйство	0,08	27500	5,7	25170	-98,6%	+9,3%

В данных, собранных в 2017 году, среднее количество публикуемых за неделю вакансий во всех областях не превысило 1, в отличие от рассчитанных нами данных методом k-средних. Но значения заработной платы незначительно отличаются в таких областях как «Страхование», «Телекоммуникация и связь», «Туризм, гостиничное дело», «Сельское хозяйство». Для профессий в сферах «Недвижимость» и «Работа дома» в значениях показателей снова присутствуют пропуски. Если эти значения также вычислить методом k-средних, то получим, что среднее количество вакансий собранных новым способом в сфере «Недвижимость» – 3,86, в сфере «Работа дома» – 2, а средняя заработная плата – 31006 руб./месяц и 24082 руб./месяц соответственно. Таким образом, учитывая эти расчеты, можно заметить, что значения по среднему количеству публикуемых вакансий близки со значениями, рассчитанными ранее, чего нельзя сказать о значениях средней заработной платы.

Если же рассчитать значения показателей, используя другую меру близости, как например, квадрат Евклидова расстояния или процент несогласия, то это не дает видимых изменений. Расхождения в данных по-прежнему остаются весомыми. Разве что при использовании в качестве нормы близости квадрат Евклидова расстояния получаем, что в сфере «Сельское хозяйство» значение заработной платы оказалось наиболее приближенным к тому, что собрала система в 2017 году. Таким образом, при данных расчетах было определено 5 кластеров. В ходе вычислений были получены следующие результаты:

1. В среднем за неделю в сфере «Сельское хозяйство» публиковалось 3,37 вакансий, а средняя заработная плата составила 27369,71 руб./месяц.
2. В сфере «ТЭК, энергетика, добыча сырья» количество опубликованных в среднем за неделю вакансий составило 2,45, а значение средней заработной платы – 25044,19 руб./месяц.
3. В сферах «Торговля оптовая», «Недвижимость» и «Телекоммуникация и связь» в среднем за неделю было опубликовано 3,06 вакансий, а средняя заработная плата составила 12521,19 руб./месяц.
4. В сфере «Страхование» среднее количество опубликованных в неделю вакансий – 3,6, значение средней заработной платы – 33123,81 руб./месяц.
5. В сферах «Работа для студентов», «Работа дома» и «Туризм, гостиничное дело» количество опубликованных в среднем за неделю вакансий составило 1,93, а значение средней заработной платы – 20422,38 руб./месяц.
6. В сфере «Туризм, гостиничное дело» значение средней заработной платы составило 20422,38 руб./месяц.

IV. Анализ полученных результатов восстановления и развитие механизма сбора данных

В целом можно сделать вывод, что применение метода k-средних в данном случае не привело к эффективному устранению пропусков в данных. Вероятно, это связанно с тем, что метод подразумевает случайный выбор кластеров, что может быть источником погрешности. Поскольку данный метод основан на вероятностном подходе, то возможна несходимость к оптимальному решению.

В рамках борьбы за полноту данных был реализован их сбор через механизмы API, предоставляемые площадками. API используется для извлечения данных с сайта и записи их в базы данных (т.е. является интерфейсом между сайтом поиска работы и системой сбора вакансий-резюме). Система взаимодействует по различным API с разными сайтами поиска работы, такими как HH.ru, Zarplata.ru, Trudvsem.ru.

Для всех трех площадок работа с API осуществляется по протоколу HTTPS. Для HH.ru и Zarplata.ru авторизация осуществляется по протоколу OAuth2. Для сайта HH.ru API поддерживает технологию CORS для запроса данных из браузера с произвольного домена. Для отладки запросов с использованием CORS доступен специальный URL в API: /cors, который поддерживает методы HEAD, GET, POST, PUT, DELETE [8]. Для площадки Trudvsem.ru все вызовы API реализуются HTTP-методом GET, предназначенным для получения информации от сервера. Количество записей, включаемых в ответ на запрос, не превышает 10 000 шт. [9]. В случае с сайтом Zarplata.ru вызовы реализуются HTTP-методом POST [10]. Ответы на запрос в API во всех трех случаях формируется только в формате JSON. Даты форматируются в соответствии с ISO 8601: YYYY-MM-DD. Метод GET используется для запроса содержимого указанного ресурса. Согласно стандарту HTTP, запросы типа GET считаются идемпотентными, т.е. то есть многократное повторение одних и тех же запросов будет возвращать один и тот же результат. Метод HEAD аналогичен методу GET. Запрос HEAD обычно применяется для извлечения метаданных, проверки наличия ресурса (валидация URL) и чтобы узнать, не изменился ли он с момента последнего обращения. В отличие от метода GET, метод POST не считается идемпотентным, то есть многократное повторение одних и тех же запросов POST может возвращать разные результаты. PUT применяется для загрузки содержимого запроса на указанный в запросе URI. DELETE удаляет указанный ресурс.

V. Прогнозирование рынка труда на основе полученных данных в сфере ИТ

С помощью собранных с сайтов поиска работы данных структурирования этих данных, возникает массив чисел, отражающий динамику спроса и предложения на рынке труда ИТ-специалистов. Для прогнозирования рынка труда использовался пакет Statistica. Одним из способов обработки полученных зависимостей является аппарат временных рядов, который позволяет выявить возможные сезонности в колебаниях спроса/ предложения на рынке труда и построить прогноз на определенный интервал времени в будущем. Временной ряд отличается от простой выборки данных, поскольку при анализе учитывается не только статистическое разнообразие и статистические характеристики выборки, но и взаимосвязь измерений со временем [11].

Прогнозная модель строится для данных собранных

системой по Новосибирской области для сферы «ИТ и интернет» на год вперед. Имеется набор данных с мая 2013 года по май 2018 (таблица 3). В таблице отражены среднее значение предлагаемой на рынке заработной платы и количество публикуемых ежемесячно вакансий в рассматриваемой сфере. В ходе анализа было выявлено что временные ряды переменных «Средняя з/п» и «Количество публикуемых вакансий» являются стационарными и имеют сезонности (рис. 5-6). Значения этих переменных возрастают каждые полгода.

Таб. 3. Количество вакансий и значение заработной платы в сфере «ИТ и интернет» в период с 2013 по 2018 год

Месяц-год	Средняя заработная плата, руб.	Количество вакансий, шт.
май-2013	33681	254
июн-2013	33154	243
июл-2013	32873	235
авг-2013	31991	214
сен-2013	36474	333
окт-2013	35942	294
ноя-2013	35682	288
дек-2013	35214	287
янв-2014	34965	226
фев-2014	34719	208
мар-2014	36671	371
апр-2014	35444	297
май-2014	34395	251
июн-2014	34368	235
июл-2014	33161	214
авг-2014	30976	207
сен-2014	36924	368
окт-2014	35789	286
ноя-2014	35712	277
дек-2014	34659	249
янв-2015	33580	231
фев-2015	31778	204
мар-2015	35587	335
апр-2015	34739	289
май-2015	33916	258
июн-2015	33865	221
июл-2015	32484	232
авг-2015	30343	221
сен-2015	35838	349
окт-2015	34744	294
ноя-2015	33875	289
дек-2015	29364	257
янв-2016	27627	255
фев-2016	27135	247
мар-2016	38690	338
апр-2016	37560	292
май-2016	36530	273
июн-2016	35240	275
июл-2016	33961	269
авг-2016	33827	259
сен-2016	37710	323
окт-2016	35360	262
ноя-2016	34941	268
дек-2016	31383	236
янв-2017	30828	244
фев-2017	30328	217
мар-2017	37349	359
апр-2017	36671	267
май-2017	35988	249
июн-2017	33863	233
июл-2017	31478	218
авг-2017	29241	204
сен-2017	33956	338
окт-2017	31222	298
ноя-2017	30514	286
дек-2017	30479	275
янв-2018	29929	267
фев-2018	28952	216
мар-2018	32648	348
апр-2018	31993	245
май-2018	30112	157

Для прогнозирования средней заработной платы и количества, публикуемых в месяц вакансий использовалась АРПСС модель скользящего среднего с параметром 1. Ошибка прогнозирования при этом составила 14,4% для значения средней заработной платы, и 12,5% для количества публикуемых вакансий.

На рисунках 7 и 8 представлены график прогноза значений средней зарплаты и график прогноза количества публикуемых вакансий соответственно.
На графиках синим цветом выделен наблюдаемый (исходный) ряд, красным – спрогнозированное значение на год вперед, зеленым – верхние и нижние границы прогноза.

Прогнозирование на основе построенной модели предполагает, что сохраняются ранее существовавшие сезонности переменных и на период прогноза. На данный момент в системе сбора данных по рынку труда сформировались реальные значения показателей за июнь и июль 2018 года. За июнь значение средней заработной платы составляет 29142 рублей, а количество вакансий – 271, относительное отклонение от значений прогноза при этом для заработной платы – 2%, а для количества вакансий – 10%. За июль значение средней заработной платы составляет 30794 рубля, а количество вакансий – 237, относительное отклонение от значений прогноза при этом для заработной платы – 10%, а для количества вакансий – 5%. Таким образом, можно сделать вывод о том, что построенная прогнозная модель эффективна.

С каждым годом в сфере информационных технологий появляются новые направления, поэтому спрос на квалифицированных ИТ-специалистов будет только расти. Предложенный метод требует небольшое количество входных данных и дает более точный результат при краткосрочном прогнозировании спроса на рынке труда, тенденции развития которого в области ИТ являются индикатором состояния рынка современных технологий.

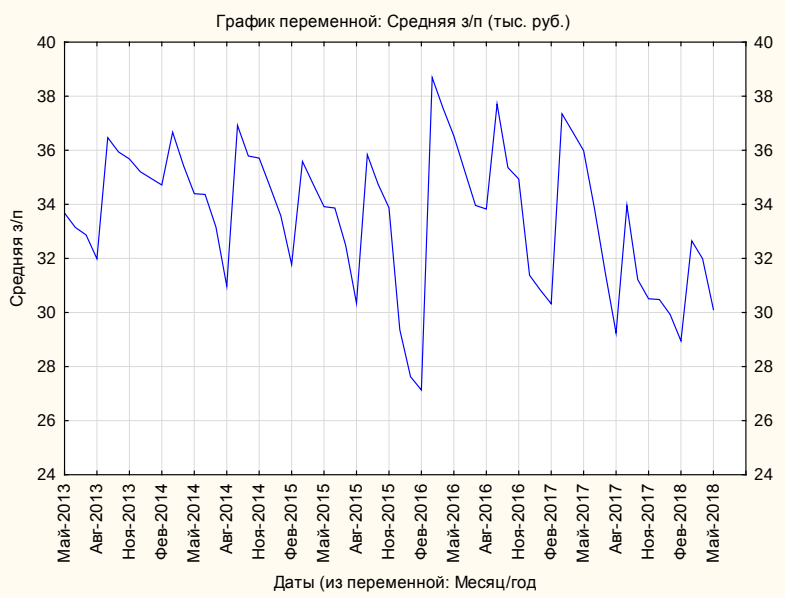

Рис. 5. График переменной «Средняя заработная плата» (в тыс. руб.)

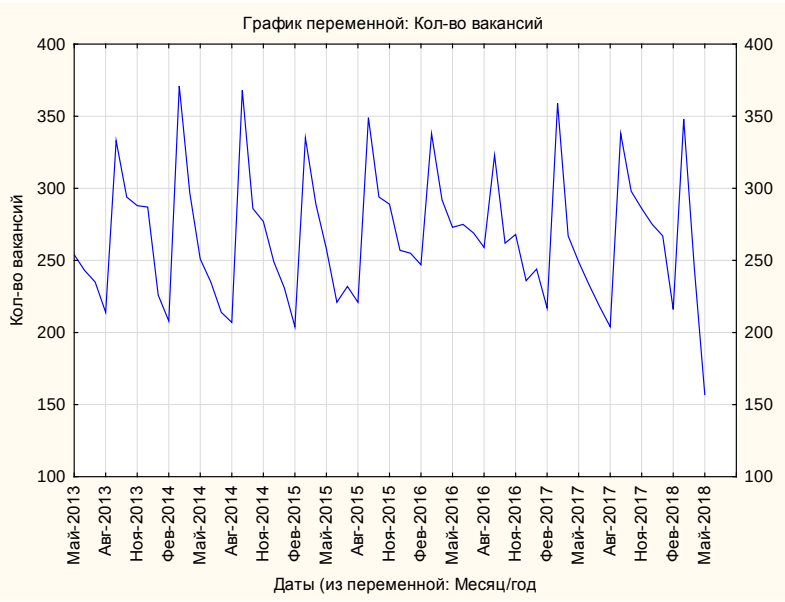

Рис. 6. График переменной «Количество вакансий»

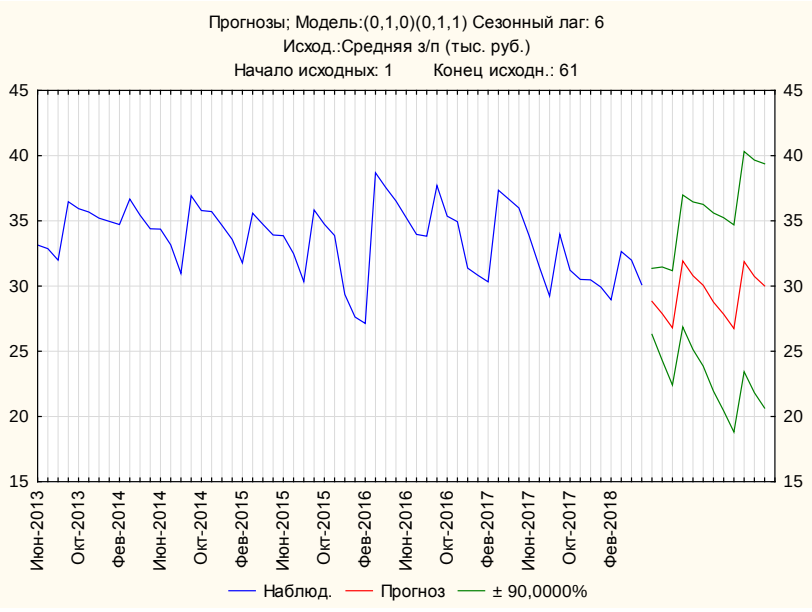

Рис. 7. График исследуемого и прогнозируемого ряда для средней заработной платы в ИТ-сфере на год вперед

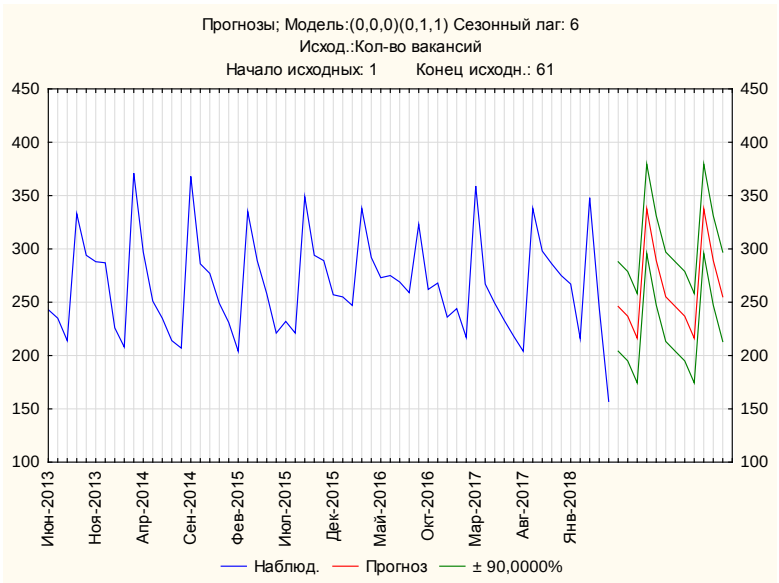

Рис. 8. График исследуемого и прогнозируемого ряда для количества публикуемых вакансий в ИТ-сфере на год вперед

VI. Выводы

В данной статье был представлен обзор наиболее распространенных методов восстановления пропусков в данных и осуществлено прогнозирование поведения рынка труда. Выбор метода устранения неполноты данных может зависеть от типов признаков, в которых имеются пропуски, от количества объектов, которые имеют пропущенные значения, а также от причины их возникновения. Применение эффективных методов позволило бы, с одной стороны, полностью устранить проблему отсутствия данных, а с другой – повысить точность прогнозируемых значений показателя. Основная идея при поиске решений заключалась в предположении о возможности спрогнозировать недостающую информацию, имея частичные данные в заданном периоде. В каждом случае необходим индивидуальный подбор метода обработки пропущенных значений. В рамках данной статьи для восстановления пропусков был использован на практике метод k-средних, реализуемый в пакете Statistica. В результате кластеризации удалось рассчитать значения для отсутствующих данных и заполнить все строки исходной таблицы по рынку труда для Новосибирской области. Но в ходе сравнения с новыми собранными данными удалось определить, что данный метод не является эффективным для заполнения пропусков. Для того чтобы получать наиболее точные данные, был предложен вариант сбора данных из API.

Построение прогнозов значений средней заработной платы и количества вакансий в сфере информационных технологий по Новосибирской области на 12 месяцев показал, что значимых изменений в направлении динамики данных показателей не предвидится. Сравнение прогнозируемых значений с уже полученными реальными данными за первые 2 месяца прогнозного периода показало, что с небольшой погрешностью модель оказалась эффективной. Прогноз дал основание полагать, что данные показатели будут возрастать с той же сезонностью, что и в предыдущих периодах.

Прогнозирование регионального спроса на рабочую силу и предлагаемую заработную плату позволяет получить интервальные оценки наиболее вероятных прогнозных значений этих показателей. Предложенный в статье метод краткосрочного прогнозирования дает теоретическое обоснование результатов прогнозирования при помощи моделей АРПСС. Как показал результат, объединение моделей авторегрессии и скользящего среднего в одной модели позволяет прогнозировать с удовлетворительной степенью точности.

Библиография

[1] Barcaroli G. Nurra A. Salamone S. Internet as Data Source in the Istat Survey on ICT in Enterprises // Austrian Journal of Statistics. 2015. Vol. 44. P. 31-43.

[2] Bakaev M., Avdeenko T. Data Extraction for Decision-Support Systems: Application in Labour Market Monitoring and Analysis // International Journal of e-Education, e-Business, e-Management and e-Learning. 2014. Vol. 4, № 1. P. 23-27.

[3] Bakaev M., Avdeenko T. Prospects and challenges in online data mining: experiences of three-year labour market monitoring project // Lecture Notes in Computer Science. 2016. Vol. 9714: Data Mining and Big Data. P. 15–23.

[4] Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies, 2002. Vol. 6, № 1. P. 51–61.

[5] Снитюк В.Е. Эволюционный метод восстановления пропусков в данных // VI МК “Интеллектуальный анализ информации”. Сб. труд. – Киев, 2006. С. 262-271.

[6] MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967. Vol. 1. P. 281–297.

[7] Боровиков В.П. Популярное введение в современный анализ данных в системе STATISTICA / Горячая линия-Телеком. М., 2013. — 288 с.

[8] Zarplata.ru API. URL: http://api.zp.ru/v1/ (дата обращения: 15.02.2018).

[9] HeadHunter API. URL: https://github.com/hhru/api/blob/master/docs/general.md (дата обращения: 15.02.2018).

[10] Открытые данные, API // Работа в России. Общероссийская база вакансий. URL: https://trudvsem.ru/opendata/apidesc (дата обращения: 15.02.2018).

[11] StatSoft. Электронный учебник по статистике. URL: http://statsoft.ru/home/textbook/default.htm (дата обращения: 27.04.2018).

Analysis and Forecasting for Labor Markets Based on Online Data

Victoria S. Giorgashvili, Maxim A. Bakaev

***Abstract* — The problem of incomplete data is quite typical in sociological, economics or statistical studies that employ online data. The possible reasons for the incompleteness are: errors and changes at the data source websites, failures and errors in the instruments for collecting data, etc. Since missing data is generally undesirable in labor market forecasting, the preferred solution is filling-in the gaps through the use of an appropriate method that wouldn't bias the results. In our paper we present a brief review of methods for eliminating incompleteness of data and describe the application of the k-means method to fill the gaps in the labor market online data that we previously collected with a dedicated software system. We evaluate the effectiveness of the method by comparing the produced results (average wages and number of ads posted by the companies) with the data additionally collected by the system through the enhanced API-based mechanism. Further, we use autoregressive integrated moving average (ARIMA) model to provide forecasts for the labor market demand in IT specialists. Validation with the data subsequently collected for the last months of 2018 suggest reasonable accuracy of the model, which can be useful in labor market monitoring and management.**

***Keywords* — data quality, missing data, web scraping, labor market, k-mean, prediction models**

REFERENCES

[1] Barcaroli G. Nurra A. Salamone S. Internet as Data Source in the Istat Survey on ICT in Enterprises // Austrian Journal of Statistics. 2015. Vol. 44. P. 31-43.

[2] Bakaev M., Avdeenko T. Data Extraction for Decision-Support Systems: Application in Labour Market Monitoring and Analysis // International Journal of e-Education, e-Business, e-Management and e-Learning. 2014. Vol. 4, # 1. P. 23-27.

[3] Bakaev M., Avdeenko T. Prospects and challenges in online data mining: experiences of three-year labour market monitoring project // Lecture Notes in Computer Science. 2016. Vol. 9714: Data Mining and Big Data. P. 15–23.

[4] Zloba E., Jackiv I. Statisticheskie metody vosstanovlenija propushhennyh dannyh // Computer Modelling & New Technologies, 2002. Vol. 6, # 1. P. 51–61.

[5] Snitjuk V.E. Jevoljucionnyj metod vosstanovlenija propuskov v dannyh // VI MK "Intellektual'nyj analiz informacii". Sb. trud. – Kiev, 2006. C. 262-271.

[6] MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967. Vol. 1. P. 281–297.

[7] Borovikov V.P. Populjarnoe vvedenie v sovremennyj analiz dannyh v sisteme STATISTICA / Gorjachaja linija-Telekom. M., 2013. — 288 s.

[8] Zarplata.ru API. URL: http://api.zp.ru/v1/ (data obrashhenija: 15.02.2018).

[9] HeadHunter API. URL: https://github.com/hhru/api/blob/master/docs/general.md (data obrashhenija: 15.02.2018).

[10] Otkrytye dannye, API // Rabota v Rossii. Obshherossijskaja baza vakansij. URL: https://trudvsem.ru/opendata/apidesc (data obrashhenija: 15.02.2018).

[11] StatSoft. Jelektronnyj uchebnik po statistike. URL: http://statsoft.ru/home/textbook/default.htm (data obrashhenija: 27.04.2018).