УДК 330.101.541

АНАЛИЗ И ПРОГНОЗ ПРЕСТУПНОСТИ В РЕГИОНАХ НА ОСНОВЕ ПОИСКОВЫХ ЗАПРОСОВ В ИНТЕРНЕТ*

А.В. Болдырева[1,2], М.А. Александров[2,3],
А.А. Кошулько[4], О.А. Соболевский[1]

[1] *Московский физико-технический институт (государственный университет)*;
[2] *Российская академия народного хозяйства и государственной службы при Президенте РФ*;
[3] *Автономный Университет Барселоны*;
[4] *Институт Кибернетики НАНУ им. В.М. Глушкова*

anna.boldyreva@phystech.edu, malexandrov@mail.ru,
koshulko@gmail.com, darina10@mail.ru

У статті розглядається застосування пошукових запитів в Інтернет для аналізу діяльності регіональної поліції та прогнозування злочинів у регіонах. Розв'язок першої задачі базується на порівнянні відносної кількості злочинів і відповідних запитів у різних регіонах. При розв'язанні другої задачі використовується наявність кореляції між динамікою враховуваних злочинів і пов'язаних з ними запитів для побудови моделей з використанням МГУА.

Ключові слова: пошукові запити, регіональна поліція, прогноз злочинів, МГУА.

The paper considers the application of web search queries for analysis of the regional police activity and forecast of crimes in regions. Solution of the first task is based on comparison of the relative number of crimes and corresponding queries in the different regions. When solving the second task, the presence of correlation between the dynamics of crimes and the queries related to these crimes is used to build models using GMDH.

Keywords: search queries, regional police, forecast of crimes, GMDH.

В статье рассматривается применение поисковых запросов в Интернет для анализа деятельности региональной полиции и прогноза преступлений в регионах. Решение первой задачи базируется на сравнении относительного количества преступлений и соответствующих запросов в различных регионах. При решении второй задачи используется наличие корреляция между динамикой учитываемых преступлений и связанных с ними запросов для построения моделей с прменением МГУА.

Ключевые слова: поисковые запросы, региональная полиция, прогноз преступлений, МГУА.

1. Введение

1.1 Постановка проблем

Имеется устойчивая статистическая зависимость между интенсивностью поисковых запросов, а также появлением и развитием различных событий и процессов, связанных с действиями информационного общества. В нашем исследовании мы используем эту зависимость для анализа и прогноза

* Эта работа есть расширенная версия статьи Boldyreva A., Alexandrov M., Koshulko O., Sobolevskiy O.: *Queries to Internet as a tool for analysis of the regional police work and forecast of the crimes in regions*, In: Proc. of 14th Mexican Int. Conf. on Artif. Intell. (MICAI-2016), Springer, LNAI, vol. 10061-10062, 2016, 12 pp.

преступлений в регионах России. Основная гипотеза здесь состоит в следующем. Пусть имеется некоторая личность, которая планирует совершить преступление или уже его совершила. Он или она пытается найти в Интернете описания каких-либо подобных случаев и/или каких-либо сопутствующих преступлению обстоятельств, информацию о наказании за преступление и о различных способах избежать его. Динамика таких запросов, количество преступлений, зарегистрированных в Генеральной Прокуратуре и данные о населении могут использоваться как для оценки деятельности региональных правоохранительных органов, так и для прогноза преступлений в регионах.

1.2 Публикации по теме

К настоящему времени число публикаций по рассматриваемым вопросам весьма ограничено, причем большинство из них относится к задаче прогнозирования преступлений, но не к задаче анализа ситуаций с преступностью. Наше исследование является попыткой привлечь внимание профессионалов к возможностям обработки поисковых запросов в Интернет для решения обеих упомянутых задач.

В [1] авторы рассматривают общие вопросы значимости и надежности, связанные с обработкой запросов поисковыми системами Интернета и звонков в сервисы мобильной связи. Среди примеров, представленных в статье, есть динамика преступлений относительно кражи и грабежа. Основной метод, применяемый авторами для прогноза, – традиционная регрессия.

Статья [2] представляет исследование на основе пространственно-временных данных Твиттера, связанные с 25 типами преступлений. Предложенный метод обработки сообщений в Твиттере позволил улучшить прогноз по 19 типам преступлений из рассматриваемых 25.

Авторы [3] описывают обощенную пространственно-временную комбинированную модель для прогноза преступлений. Они объединяют запросы в поисковой системе Интернет и посты в Твиттере, чтобы улучшить качество их предыдущей модели.

Техника Hotspot Mapping для прогнозирования пространственных характеристик преступлений представлена в [4]. Авторы рассматривают 4 типа преступлений: кражу, уличные преступления, угон транспортных средств и кражу транспортных средств, и показывают, как настроить упомянутую технику на разные виды преступлений.

1.3 Вклад авторов работы

Наша первая статья в этой области относится к прогнозу экономических преступлений с применением 2-х методов: традиционный регрессионный анализ (РА) и Метод Группового Учета Аргументов (МГУА) [5]. Здесь зависимая переменная отражает общее количество преступлений,

принадлежащих к категории экономических, применительно ко всей территории России, а независимые переменные – это поисковые запросы в Интернет. Мы выбрали запросы, отражающие содержание упомянутых преступлений и имеющие значительную корреляцию с их динамикой. Эксперименты продемонстрировали значительное преимущество МГУА над РА, поэтому в наших дальнейших исследованиях, связанных с задачами прогноза, мы использовали только этот метод – в частности, платформу GMDH Shell, содержащую модифицированные алгоритмы МГУА [6].

В [7] мы изучали различные алгоритмы МГУА и разные процедуры предварительной обработки данных для предсказания экономических преступлений. В этой работе мы использовали так называемые «барометры», являющиеся композицией 10 временных рядов поисковых запросов, имеющих наиболее сильную корреляцию с динамикой упомянутых преступлений. Эти барометры представляли собой модификацию тех, которые компания Google использовала в одном из ее ранних приложений [8]. Здесь термин «алгоритмы МГУА» означает разные варианты комбинаторных и нейросетевых алгоритмов на платформе GMDH Shell, а термин «предварительная обработка» означает различные преобразования данных – например, логарифм, квадратный корень и так далее. Результаты оказались очень многообещающими: достигнутый уровень MAPE был равен 2-6%, где MAPE обозначает известную среднюю абсолютную ошибку в процентах (Mean Absolute Percentage Error).

Отличие нынешнего исследования от наших предыдущих работ:

- впервые выполняется анализ данных о работе региональной полиции;
- выполняется прогноз вполне определенных преступлений в конкретных регионах в отличие от предыдущих исследований, где рассматривалась целая категория преступлений по всей территории страны;
- используется 153 дескриптора вместо 10-20, с которыми мы имели дело в наших экспериментах из предыдущей статьи.

2. Данные

2.1 Индикаторы и дескрипторы

Здесь и далее мы используем следующие обозначения:

- SQ (Search Query), поисковые запросы в Интернет;
- RC (Registered Crime), зарегистрированные преступления в Генеральной Прокуратуре России;
- УК, Уголовный кодекс России;
- FD (Federal District), федеральные округа России.

SQ и RC учитываются в относительных единицах: SQ нормализуется на сумму поисковых запросов в данном регионе, а RC – на общее количество населения в данном регионе.

Статистические данные по преступленям в регионах России находятся в открытом доступе на сайте Прокуратуры [9]. В нашем исследовании мы рассматриваем преступления, отраженные в следующих 3-х статьях УК:

- Статья 111 УК – Намеренное причинение тяжких телесных повреждений;
- Статья 222 УК – Незаконная покупка, передача, продажа, хранение, транспортировка оружия;
- Статья 291 УК – Взяточничество.

Рассматриваемый отрезок времени покрывает чуть более 2 лет с октября 2013 по декабрь 2015. Такое ограничение определяется условиями компании Yandex [10]. Мы предпочитаем использовать статистику запросов именно компании Yandex, поскольку ее данные представлены и в абсолютных, и в относительных единицах. Google же представляет свои данные только в абсолютной форме [11].

Мы создали базу данных (БД) дескрипторов, которые покрывали эти статьи УК, дав им простые названия: «статья 111», «статья 112», «статья 113», и т.д. Кроме того, мы включили в БД ряд словосочетаний, которые могут быть полезны для прогнозирования: «незаконный въезд», «покупка оружия», «ограничение свободы», «наказание», «особенно большой», «особенно серьезный», «ухудшение», «нарушение», «использование оружия» и др.

2.2 Гипотезы

Выше мы предположили, что человек, который планирует совершить преступление или уже сделал это, обращается в Интернет, чтобы найти любую полезную информацию относительно такого рода преступлений. Этот человек интересуется наказанием за преступление, для чего он или она использует SQ с прямым упоминанием соответствующей статьи (статей). При таких предположениях мы можем зафиксировать следующие конкретные гипотезы:

1. Число людей, совершивших данное преступление в течение данного срока в данном регионе, является определенной долей всех SQ, отражающих это преступление
2. Количество SQ с прямым упоминанием о статье (статьях), связанной с данным преступлением, является определенной долей всех SQ, отражающих это преступление

Таким образом, в рамках этих гипотез за SQ «статья X» стоит определенное число реальных людей. С этим допущением мы можем интерпретировать результаты, представленные далее в разделах 3 и 4.

При этом некоторые простые заключения могут быть сделаны без какой-либо дополнительной обработки. Например, пусть динамика индикатора «взятка» имеет сильную корреляцию с динамикой дескрипторов «статья 191» (незаконная торговля драгоценными металлами, натуральными драгоценными камнями или жемчугом) и «статья 269» (нарушение правил техники безопасности во время строительства, эксплуатации и обслуживания главных

трубопроводов). В этом случае можно сделать предположение об определенных злоупотреблениях с материалами, связанными со строительством.

3. Анализ деятельности региональных правоохранителей

3.1 Тренд-анализ SQ и RC

Мы можем сформулировать следующие утверждения, связанные с анализом преступности в регионах:

- Согласно гипотезам, представленным в п. 2.2, чем большее различие между SQ и RC, тем более проблемной является ситуация в данном регионе относительно данной статьи УК;
- Чтобы исключить зависимость от количества населения в каждом регионе, RC должно быть представлено в относительном выражении;
- Чтобы учесть Интернет-активность в данном регионе, переменная SQ должна быть представлена в относительном выражении.

Наш подход к оценке работы региональной полиции в данном регионе с данным типом преступлений состоит в двух идеях:

(1) сравнение SQ и RC в рассматриваемом регионе;

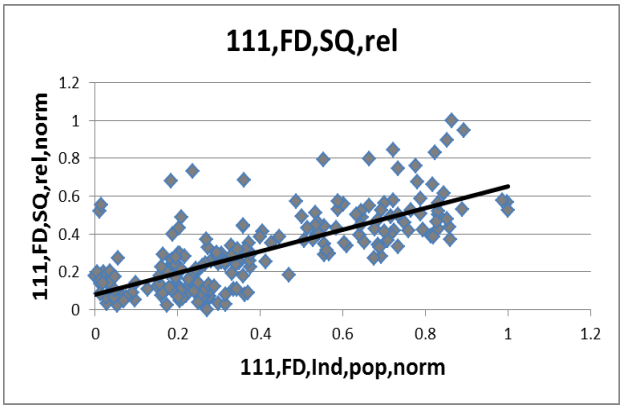
(2) сравнение ситуации данного региона с другими регионами.

Чтобы правильно интерпретировать результаты таких сравнений, мы должны прежде удостовериться в положительной связи количества SQ и числа RC. Для этого в нашем начальном эксперименте мы рассмотрели связи между SQ «статья 111», «статья 291» и «статья 222» и официальной статистикой преступлений по этим статьям [9].

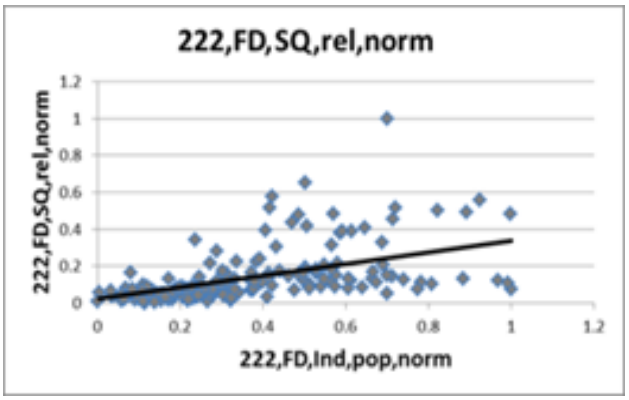
Статистика преступлений принимает во внимание количество населения в регионе. Результаты представлены на Рис. 1, где каждый регион (федеральный округ) отражен одним значком. Период времени был с октября 2013 по декабрь 2015, что составляет 27 месяцев, поэтому каждая диаграмма содержит 27x8 = 216 значков. Данные здесь пронормированы таким образом: вначале мы нормировали число запросов и число зарегестрированных преступлений по статьям в каждом регионе на население региона, а затем полученные значения были промасштабированы на интервалы [0, 1] как для поисковых запросов, так и для зарегестрированных преступлений. Результаты, представленные на Рис.1, оказались не совсем удобными для восприятия и интерпретации, поскольку положение некоторых регионов было довольно неожиданным.

Для более удобного визуального представления числа RC были повторно промасштабированы на 100,000, 100,000 и 1,000,000 населения для статей 111, 222 и 291 соответственно. Кроме того, числа SQ были разделены на 1,000, 1,000 и 100 для тех же статей 111, 222 и 291, соответственно. Результаты представлены на Рис. 2.

(а)



(б)



(в)

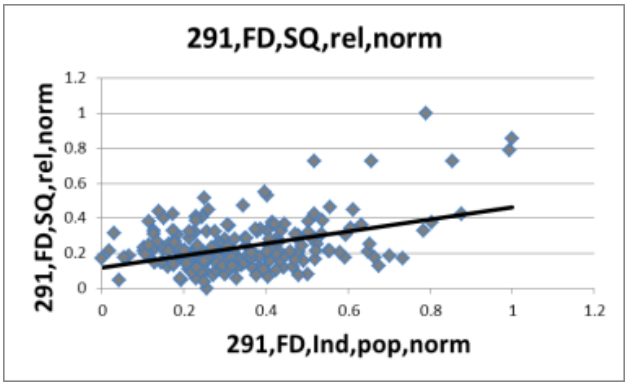
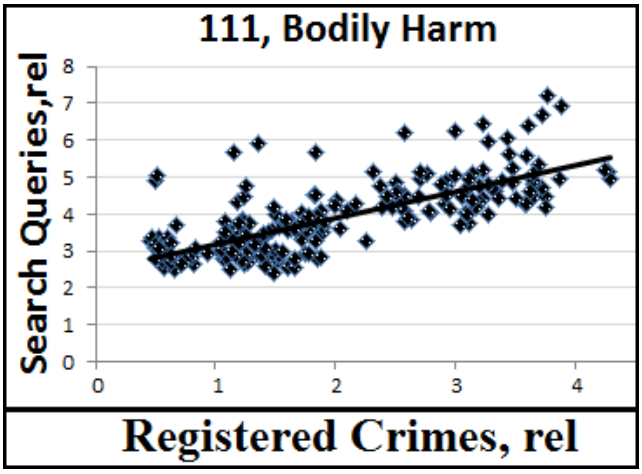


Рис. 1. Зависимость (RC,SQ) для 8 федеральных округов,
(а) для статьи 111, (б) для статьи 222, (в) для статьи (291),
масштабирование на [0,1]

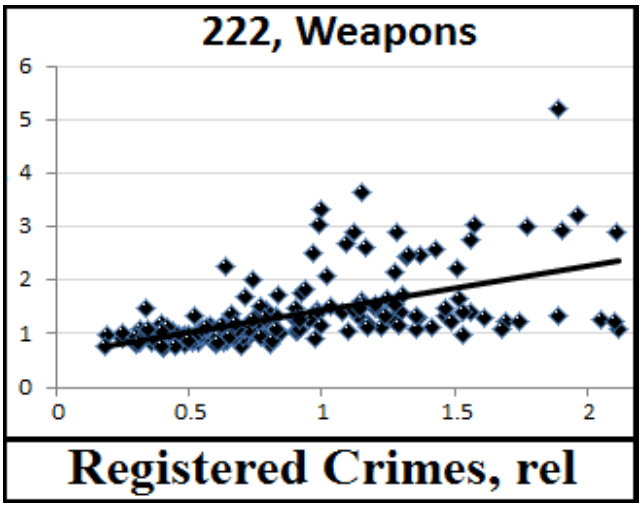
Тренд-анализ демонстрирует, что количество поисковых запросов в регионах увеличивается с ростом зарегестрированных преступлений. Кроме того, легко видеть, что тенденция роста намного сильнее выражена для статьи 111 (преднамеренное причинение тяжких телесных повреждений) по сравнению со статьями 222 (незаконные операции с оружием) и 291 (взяточничество). Такая ситуация может быть легко объяснена следующим: в отличие от статьи 111, в статьях 222 и 291 нет лиц, заинтересованных в

раскрытии этих преступлений. Поэтому много преступлений, связанных с этими статьями, остаются не зарегистрированными.

(а)



(б)



(в)

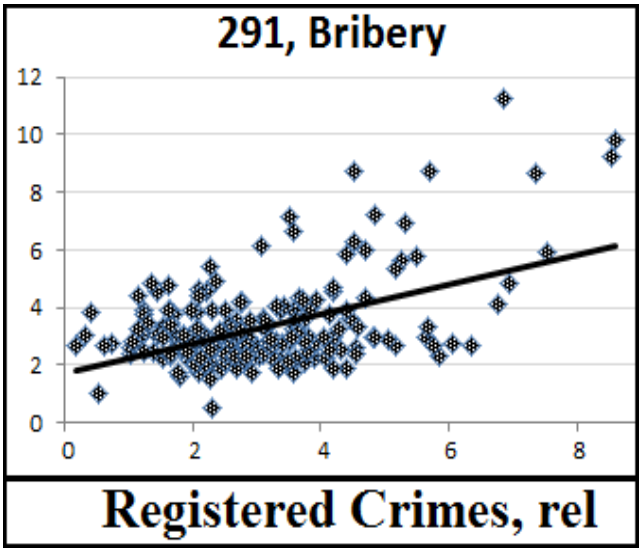


Рис. 2. Зависимость (RC,SQ) для 8 федеральных округов,
(а) для статьи 111, (б) для статьи 222, (в) для статьи (291)

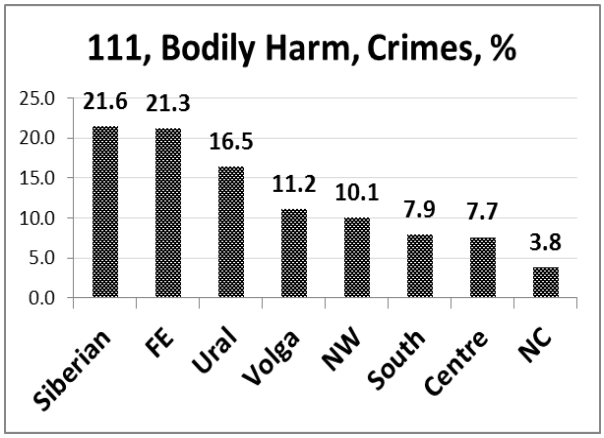
3.2 Метод сравнения масштабированных RC и SQ

Мы провели анализ RC и SQ, относящихся к 3-м рассматриваемым статям, в 8 федеральных округах, для чего выполнили следующие процедуры предварительной обработки:

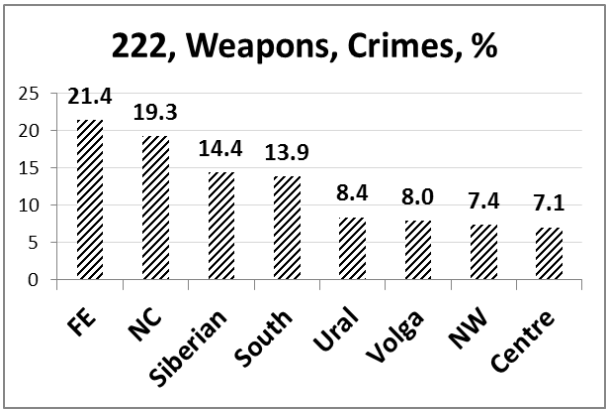
- RC данные каждого региона были нормированы на общую численность населения этого региона. Затем эти значения были нормализованы на их сумму для приведения к единичному интервалу;
- SQ данные каждого региона были нормализованы на их сумму.

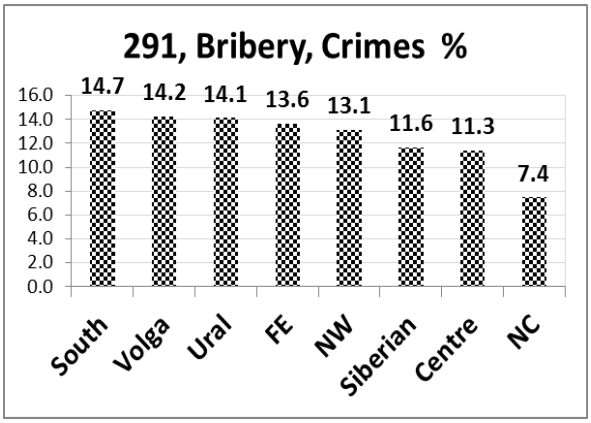
Результаты распределения RC и SQ представлены на Рис. 3 и 4 (в процентной форме), где FE (Far East), NW (North West) и NC (North Caucasus) обозначают Дальний Восток, Северо-Запад и Северный Кавказ соответственно.

(а)



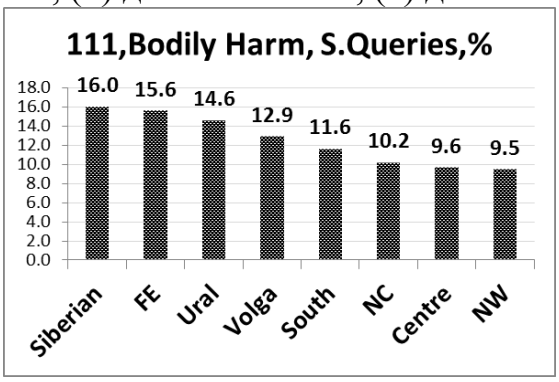
(б)



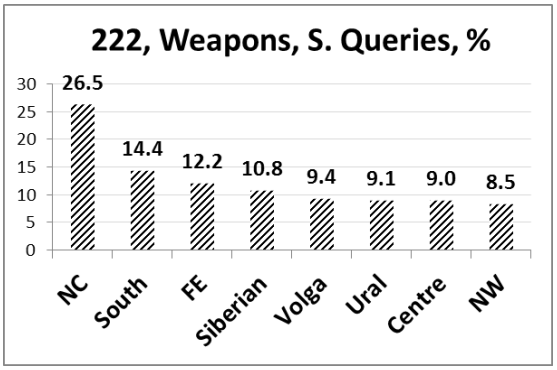


(в)

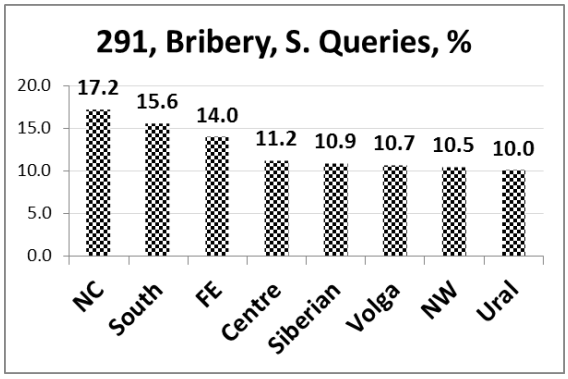
Рис. 3. Зарегестрированные преступления (RC) в относительных единицах для федеральных округов как процент от их общего числа; (а) для статьи 111, (б) для статьи 222, (в) для статьи (291)



(а)



(б)



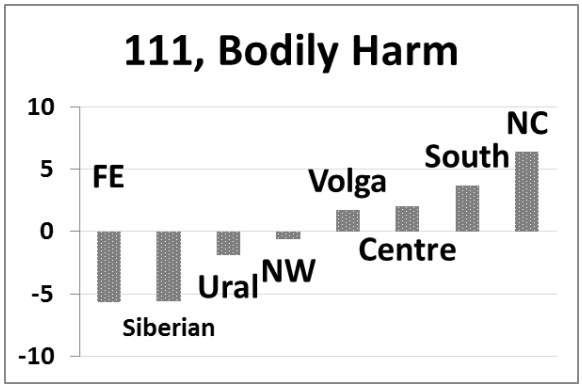
(в)

Рис. 4. Поисковые запросы (SQ) в относительных терминах в федеральных округах как процент от их общего числа: (а) для статьи 111; (б) для статьи 222; (в) для статьи (291)

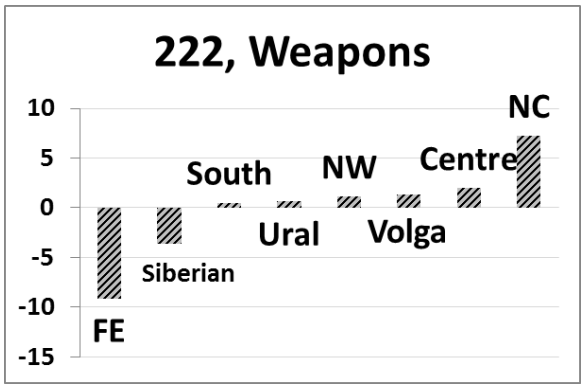
Чтобы оценить деятельность региональной полиции, нужно рассмотреть различие между расчитанными относительными значениями RC и SQ. Эти результаты представлены на Рисунке 5. Регионы на рисунке упорядочены согласно этому различию.

Левые части гистограмм показывают регионы, где уровень RC превышает уровень SQ. Здесь, таким образом, мы можем предположить хорошую работу региональной полиции. И наоборот, правые части гистограмм показывают регионы, где уровень RC меньше, чем уровень SQ. И здесь мы можем говорить о плохой работе полиции. Очевидно, что понятия «хороший» и «плохой» имеют здесь относительный смысл.

(а)



(б)



(в)

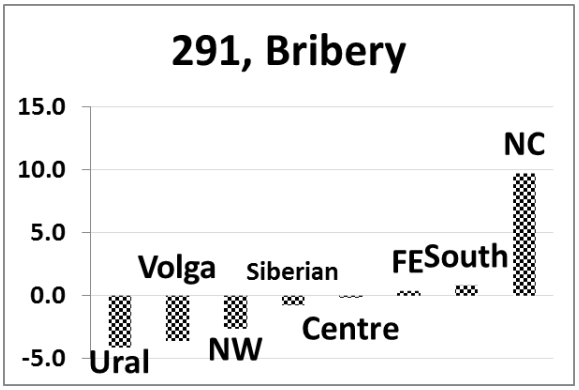


Рис. 5. Различие между зарегестрированными преступлениями (RC %) и поисковыми запросами (SQ %):
(а) для статьи 111; (б) для статьи 222; (в) для статьи (291)

Для подтверждения результататов нашего анализа нужна была бы официальная информация от российского МВД – однако эта информация является закрытой, и сайт Министерства ничего не говорит об оценках работы региональной полиции [12]. Единственным источником информации для нас были российские СМИ, где утверждается, что дальневосточные, сибирские и северо-западные регионы демонстрируют наилучшую ситуацию борьбы с преступностью. Худшим регионом по всем видам преступлений является Северный Кавказ. В целом наши результаты согласуются с этими оценками, отраженными в СМИ.

4. Прогноз преступлений

4.1 Типы моделей

Мы строим 3 модели для прогнозирования преступлений, связанных с упомянутыми выше статьями 111, 222 и 291. Горизонт прогноза составляет 3 месяца. Согласно мнению следователей полиции, время между планированием преступления и его совершением обычно не превышает 3 месяца. Именно поэтому мы рассматриваем 3 месяца как ограничение возможного лага в моделях, и используем построенные модели, чтобы предсказать преступления с упреждением 1, 2 и 3 месяца.

Авторегрессионная модель. Самая простая модель, для которой не нужна никакая дополнительная информация, это модель авторегрессии. Здесь предполагается, что все свойства заданного временного ряда полностью представлены в предыстории процесса. Такие модели являются самыми популярными среди экспертов, и в обобщенном виде модель этого типа может быть представлена в форме:

$$y_{t+1} = F(y_t, y_{t-1}, y_{t-2}, y_{t-3}).$$

Здесь y_i – значения рассматриваемого индикатора (статьи 111, 222, 291) в момент i, t – текущий момент времени.

Регрессионная модель. Следующий тип моделей – это регрессия на поисковых запросах, и эти модели являются для нас самыми интересными. Основной проблемой, с которой мы здесь столкнулись, было огромное потенциальное количество регрессоров SQ, претендующих на включение в регрессионную модель. Чтобы обойти эту трудность, мы использовали так называемые барометры, являющиеся усредненными временными рядами тех SQ, которые имеют наибольшую корреляцию с заданным индикатором для нескольких лагов. Барометры создаются отдельно для каждого типа преступлений. Для этого необходимо выполнить две процедуры:

(1) Формирование топ-выборок с высокой корреляцией значений SQ с индикатором. Говоря «корреляция», мы имеем здесь в виду и положительную, и отрицательную корреляцию. В наших экспериментах

мы использовали порог корреляции 0.7 как компромиссный, поскольку с порогом 0.5 список временных рядов SQ содержит несколько сотен, что слишком много, а с порогом 0.9 этот список уменьшается до одного-двух десятков выборок, вследствие чего мы теряем информацию. Уровень корреляции проверяется по критерию Пирсона для текущего момента и для лагов в 1, 2, и 3 месяца. Всего было отобрано 153 переменных SQ.

(2) Создание 8 барометров, из которых 4 содержат средние значения временных рядов топ-выборок, имеющих самую высокую корреляцию с индикатором для лагов 0, 1, 2, 3 месяца соответственно, а другие 4 барометра – средние значения временных рядов топ-выборок, имеющих высокую антикорреляцию с индикатором для тех же самых лагов.

Получающаяся обобщенная модель может быть представлена в форме:

$$y_{t+1} = F(b_{1,t},\ldots b_{1,t-3},b_{2,t},\ldots,b_{2,t-3},\ldots,b_{8,t},\ldots,b_{8,t-3}),$$

где $b_{i,j}$ – барометр, i – его номер, j – момент времени.

Комбинированная модель. Этот тип моделей включает как переменные авторегрессии, так и лаговые регрессоры:

$$y_{t+1} = F(b_{1,t},\ldots,b_{1,t-3},b_{2,t},\ldots,b_{2,t-3},\ldots,b_{8,t},\ldots,b_{8,t-3}, y_t, y_{t-1},y_{t-2},y_{t-3}).$$

Все обозначения здесь были описаны выше.

3.2 Метод моделирования

***Метод Группового Учета Аргументов* (*МГУА*).** Чтобы построить конкретную модель, мы используем МГУА. Этот метод был предложен украинским ученым академиком А.Г. Ивахненко в 70-х годах прошлого века, и в настоящее время МГУА развивается его учениками и последователями. В работе [13] дан подробный обзор традиционных и новых алгоритмов МГУА. Теоретические основы метода представлены в статье [14]. Список алгоритмов, отражающих подход МГУА, содержит десятки модификаций. Как алгоритмы, так и их многочисленные приложения регулярно представляются на ежегодных Международных конференциях и семинарах в Украине и европейских странах. Например, можно упомянуть здесь известную конференцию ICIM-2013, посвященную памяти А.Г. Ивахненко [15].

МГУА – это метод (или скорее даже технология), который позволяет строить модель оптимальной сложности в заданном классе моделей. Метод демонстрирует свои преимущества, когда: а) объем экспериментальных данных очень ограничен, и б) априорная информация о модели, которая строится, отсутствует или почти отсутствует.

Метод состоит в выполнении следующих шагов:

1. Эксперт определяет класс моделей, в котором автоматически будет генерироваться достаточно большое число (десятки и сотни тысяч) моделей различной структуры от простых до самых сложных.

2. Экспериментальные данные делятся на две части: одна для обучения (оценивания параметров) моделей (обучающая выборка), и вторая для проверки качества генерируемых моделей (проверочная выборка). Такое деление проводится вручную или с использованием специальной автоматической процедуры.
3. Для текущей модели лучшие параметры определяются на обучающей выборке. Здесь используются любые внутренние критерии соответствия между моделью и данными, например, критерий наименьших квадратов.
4. Эта модель проверяется на второй части данных с использованием каких-либо внешних критериев. Например, это может быть среднеквадратичное отклонение значений выхода модели от проверочных данных.
5. Внешние критерии (или самый важный из них) проверяется на достижение устойчивого оптимума на множестве генерируемых моделей. В этом случае поиск модели заканчивается, иначе рассматривается более сложная модель, и процесс повторяется с шага 3.

Пояснение, почему внешние критерии достигают оптимума (минимума). Предполагается, что экспериментальные данные содержат: а) регулярную компоненту, определяемую структурой модели, и б) случайную компоненту — шум. Очевидно, модель должна быть способной отражать свойства регулярной компоненты. Когда модель слишком проста, она слабо отражает регулярную компоненту и нечувствительна к шуму (модель недообучена). Когда модель слишком сложна, она хорошо отражает как регулярную компоненту, так и изменения случайной компоненты (модель переобучена). В обоих случаях значения функции штрафа (внешний критерий) оказываются большими. Поэтому мы ожидаем найти точку, где критерий достигает своего минимума.

GMDH Shell как платформа для применения МГУА. GMDH Shell является платформой для моделирования с применением алгоритмов МГУА [6], и именно она использовалась в наших экспериментах. GMDH Shell включает:

- начальные преобразования исходных данных (логарифмы, квадраты, корни и др.) с лагами и дополнительными переменными;
- основной комбинаторный алгоритм (COMBI) и основной нейросетевой алгоритм (многорядный итерационный алгоритм MIA) с модификациями;
- различные способы контроля качества модели в виде различных вариантов кросс-валидации и т.д.

Оба упомянутых алгоритма описаны подробно в [13].

COMBI является именем группы алгоритмов, которые сравнивают все возможные комбинации переменных в модели в рамках заданного класса. Ниже даны примеры таких классов моделей.

А. Класс моделей – линейные функции от n переменных:

Уровень 1: $y_i = a_0 + a_i x_i \qquad i = 1,2,\ldots n$

Уровень 2: $y_k = a_0 + a_i x_i + a_j x_j \qquad i,j = 1,2,\ldots n;$

и т.д.

В. Класс моделей – полиномиальные функции от n переменных с ограничением степени (например 2):

Уровень 1: $y_k = a_0 + a_i x^p_i \qquad i = 1, 2,\ldots,n; \quad p = 1, 2$

Уровень 2: $y_k = a_0 + a_i x^p_i + a_j x^q_j \qquad i,j = 1, 2,\ldots,n; \quad p,q = 1, 2$

Уровень 3: $y_k = a_0 + a_i x^p_i + a_j x^q_j + b_{ij} x_i x_j \qquad i,j = 1, 2,\ldots,n; \quad p,q = 1, 2$

и т.д.

MIA – это имя группы алгоритмов нейросетевого типа на основе наследования лучших комбинаций переменных. Здесь фиксируется число лучших моделей на заданном слое, которые порождают новые переменные на следующем слое, используя заданную функцию преобразования:

Слой 1: $y_i = a_0 + a_i x_i,\ i = 1, 2,\ldots,n$

Слой 2: $z_k = a_0 + f(y_i,y_j),\ i,j = 1, 2,\ldots,n;$ (i,j – номера выбранных моделей)

Слой 3: $g_k = a_0 + f(z_i,z_j),\ i,j = 1, 2,\ldots,n;$ (i,j – номера выбранных моделей)

и т.д.

Типичные функции трансформации f следующие:

(1) Линейная: $f(s,t) = a_0 + a_1 s + a_2 t$

(2) Квадратичная : $f(s,t) = a_0 + a_1 s + a_2 t + a_3 st + a_4 s^2 + a_5 t^2$

3.3 Результаты моделирования

Мы изучили 10 вариантов построения моделей, используя оболочку GMDH Shell. Эксперименты отличались следующими характеристиками:

- алгоритмы – COMBI или MIA;
- переменные – авторегрессионные члены, барометры, или то и другое вместе;
- степени полинома или функции нейрона – линейные (1) или квадратичные (2).

Для проверки качества модели мы использовали процедуру *two-fold cross validation*, который в терминах МГУА носит название «симметричный критерий регулярности». Последние 3 месяца были исключены из рассмотрения и использовались для контрольной проверки точности прогноза. Таблица 1 содержит описания моделей и абсолютные процентные ошибки прогноза на 1, 2, 3 месяца (с их средними значениями) для статьи 111. Здесь 1, 2 и 3 – обозначения месяцев. Лучшая модель относительно среднего значению точности прогноза была построена с помощью алгоритма COMBI. Это отмечено в Таблице 1.

Что касается статей 222 и 291, то их ошибки прогноза на 1 месяц для лучших 3 моделей составляют 15% и 10% соответственно. Все эти модели включают барометры.

Чтобы продемонстрировать реальную динамику зарегестрированных преступлений (RC) для статей 111, 222 и 291, мы показываем 3 скриншота. Они

также отражают работу оболочки GMDH Shell. Рис. 6 относится к статье 111, прогноз был выполнен с использованием алгоритма MIA. Рис. 7 и Рис. 8 относятся к статьям 222 и 291 соответственно, а прогноз был выполнен с применением алгоритма COMBI. Оба алгоритма MIA и COMBI использовали вместе авторегрессию и барометры.

Таблица 1

Варианты моделирования и ошибки прогноза для статьи 111

Алгоритм	***Авторегр***	***Барометр***	***Степень***	**1, %**	**2, %**	**3, %**	***Среднее, %***
COMBI	+	–	1	5,0	3,3	0,9	3,1
COMBI	–	+	1	1,0	2,8	3,9	2,6
COMBI	–	+	2	1,0	6,2	3,2	3,5
COMBI	+	+	1	5,0	3,3	2,0	3,4
COMBI	+	+	**2**	**2,0**	**0,2**	**0,3**	**0,8**
MIA	+	–	1	7,9	1,3	9,5	6,2
MIA	–	+	1	1,0	2,8	0,8	1,5
MIA	–	+	2	2,4	4,8	0,8	2,7
MIA	+	+	1	1,8	8,7	1,7	4,1
MIA	+	+	2	2,0	2,2	3,7	2,6

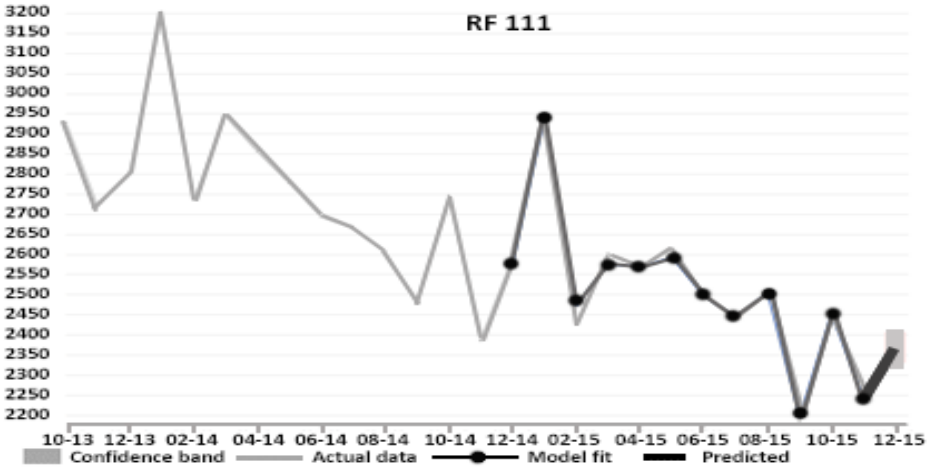


Рис. 6. Прогноз для статьи 111 с применением алгоритма MIA

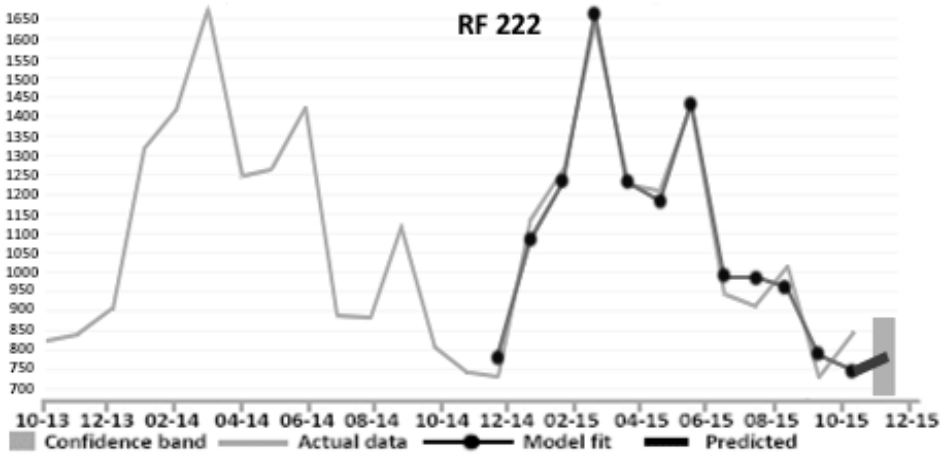


Рис. 7. Прогноз для статьи 222 с применением алгоритма COMBI

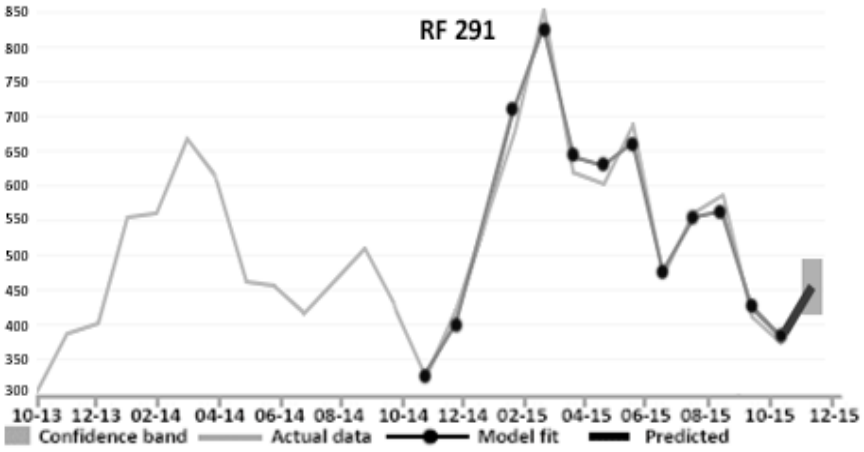


Рис. 8. Прогноз для статьи 291 с применением алгоритма COMBI

Общие результаты выполненного моделирования следующие:

- Барометры улучшают качество прогноза для всех моделей и всех статей. Только в одном случае эксперимента со статьей 111 результаты остались теми же самыми.
- Прогноз для статьи 111 значительно лучше, чем для других статей. Причина, с нашей точки зрения, состоит в том, что данные регистрации преступлений по статьям 222 и 291 по своей природе содержат определенные искажения – ведь в преступлениях по статье 111 всегда есть жертва, заинтересованная в регистрации и раскрытии преступления, в то время как в преступлениях по статьям 222 и 291 таких жертв нет.

Заключение

В данном исследовании предложена методика оценки деятельности региональных правоохранительных органов, которая состоит в сравнении количества зарегистрированных преступлений и количества поисковых запросов после специальных процедур масштабирования. В экспериментах мы рассматривали работу правоохранителей в 8 федеральных округах России относительно 3 заданных типов преступлений. Полученные результаты в целом соответствуют информации, отраженной в российских СМИ.

Мы также изучили возможность предсказывать упомянутые преступления, используя различные модели зависимости количества преступлений от интенсивности поисковых запросов в Яндекс по определенным ключевым словам. Такие модели мы построили с применением алгоритмов МГУА, имеющихся на платформе GMDH Shell, и выбрали наилучшие модели для каждого типа преступлений. Точность прогноза оказалась очень высокой для преступлений, где есть жертвы, и ниже для преступлений без жертв.

Мы полагаем, что предложенные методики могут быть полезны для региональных правоохранительных органов.

Литература

1. Zhu J., Wang X., Qin J., Wu L. Assessing Public Opinion Trends based on User Search Queries: Validity, Reliability, and Practicality. – Proc. of 65th Annual Conf. of the World Association for Public Opinion Research. – Hong Kong: PRC, 2012. – P. 1-7.
2. Gerber M.S. Predicting Crime using Twitter and Kernel Density Estimation. – Web: //www.sciencedirect.com/science/article/pii/S0167923614000268.
3. Wang X., Brown D.E., Gerber M.S. Spatio-Temporal Modeling of Criminal Incidents using Geographic, Demographic, and Twitter-derived Information. - Proc. of Int. Conf. on Intelligence and Security Informatics (ISI-2012). – P. 36-41. – Web: //ptl.sys.virginia.edu/ptl/sites/default/files/ISI2012_WBG.pdf.
4. Chaineya S., Tompson L., Uhlig S. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime // Security Journal. – London: University College, 2008. – V. 21. – P. 4–28. – Web: //www.palgrave-journals.com/sj/journal/v21/n1/full/8350066a.html, (2008).
5. Болдырева А., Кошулько А. Прогнозные модели экономических преступлений на основе поисковых запросов в Интернет: Регрессия vs МГУА / Труды социол. фак-та МГУ «Математическое моделирование социальных процессов». – М.: МГУ, 2015. – Вып.17. – С. 34-42.
6. GMDH Shell: platform for inductive modeling.–Web: //www.gmdhshell.com/.

7. Boldyreva A., Koshulko O. GMDH Helps to Build Models Based on Queries to Yandex for the Forecast of Economic Crimes / Proc. of 7th Int. Workshop on Inductive Modeling. – Kyiv: IRTC ITS NAS of Ukraine, 2015. – P. 9-11. – Web: //www.mgua.irtc.org.ua/attach/ICIM-IWIM/2015/index2015.html.
8. Espino, J., Hogan, W., Wagner, M. Telephone triage: A Timely Data Source for Surveillance of Influenza-like Diseases / Proc. AMIA Symp. – 2003. – P. 215–219.
9. Генеральная Прокуратура РФ, официальная статистика. – Web: //crimestat.ru/offenses_map/.
10. Статистика Yandex. – Web: //wordstat.yandex.com/.
11. Google тренды. Web: //www.google.ru/trends/.
12. Материалы МВД России. – Web: //xn--b1aew.xn--p1ai/.
13. Stepashko, V. Ideas of Academician O. Ivakhnenko in Inductive Modeling Field from Historical Perspective / Proc. of 4th Intern. Conf. on Inductive Modeling (ICIM-2013). – Kyiv: IRTC ITS NAS of Ukraine, 2013. – P. 31-37.
14. Stepashko, V. Method of Critical Variances as an Analytical Tool of the Inductive Modeling Theory // Journ. of Inform. and Automat. Sciences. – Begell House Inc., 2008. – V. 40. – No 3. – P. 47-58.
15. Proc. of the 4th Intern. Conf. on Inductive Modeling (ICIM-2013) / Stepashko, V. (Ed.). – Kyiv: IRTC ITS NAS of Ukraine, 2013. Web: // www.mgua.irtc.org.ua/attach/ICIM-IWIM.