

© 2015 г. **К.В. РУДАКОВ**, чл.-корр. РАН
(ФИЦ ИУ РАН, Вычислительный Центр им. Дородницына РАН, Москва),
М.П. Кузнецов,
А.П. Мотренко,
М.М. Стенина
(Московский физико-технический институт, Москва),
Д.О. Каширин,
В.В. Стрижов, д-р физ.-мат. наук,
(ФИЦ ИУ РАН, Вычислительный Центр им. Дородницына РАН, Москва)

ВЫБОР ОПТИМАЛЬНОЙ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОБЪЕМОВ ГРУЗОВЫХ ЖЕЛЕЗНОДОРОЖНЫХ ПЕРЕВОЗОК 02.50.Sk, 02.60.Ed¹

Решается задача выбора оптимальной модели краткосрочного прогнозирования объемов железнодорожных перевозок по историческим и экзогенным временным рядам. Исторические данные содержат информацию об объемах перевозок различных типов грузов между парами станций. Предполагается, что результат выбора оптимальной модели зависит от уровня агрегирования по типам грузов, пунктам отправления и назначения и по времени. Рассмотрены модели векторной авторегрессии, интегрированная модель авторегрессионного скользящего среднего и непараметрическая модель гистограммного прогнозирования. Предложены критерии сравнения прогнозов на основе расстояний между ошибками прогнозов моделей. Данные критерии используются для анализа моделей с целью определения допустимых запросов на прогноз, в том числе, фактической глубины прогнозирования.

Ключевые слова: *прогнозирование временных рядов, грузовые железнодорожные перевозки, выбор модели прогнозирования*

1. Введение

Грузовые перевозки обеспечивают свыше 80% общей выручки железнодорожного транспорта [1]. В связи с этим, прогнозирование объемов перевозок имеет большое практическое значение для их планирования с целью повышения экономической эффективности грузовых железнодорожных перевозок. Для повышения качества выполняемых прогнозов предлагается учитывать как исторические временные ряды самих грузоперевозок, так и экзогенные временные ряды, содержащие информацию о влиянии внешних факторов на объемы перевозок.

Исторические данные (см. табл. 1) содержат информацию об отправлениях грузов: дату погрузки, станцию отправления, станцию назначения, количество вагонов, которые про-

¹Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (соглашение № RFMEFI60414X0041).

шли по маршруту от станции отправления до станции назначения, код груза, род вагонов, суммарный вес груза в тоннах и признак маршрутной отправки.

Прогнозы выполняются по запросу заданного формата. Запрос на прогноз задает уровень агрегирования данных. Учитывая формат данных, возможна детализация

1) по типам грузов: по всем грузам, либо по заданному грузу. Прогнозы разрабатываются для номенклатуры грузов, учитывающей до 41 наименования грузов,

2) по пунктам отправления и назначения: парам станций или районов. Единая сетевая разметка включает 99 сетевых районов, причем каждый сетевой район включает до 100 станций, открытых для грузовых операций,

3) по времени: по дням, по неделям, по месяцам.

Значения прогнозов вычисляются посредством модели. Моделью называется [3] параметрическая функция, которая по временному ряду (или набору временных рядов одинаковой длины) и фиксированным параметрам вычисляет единственную точку – прогноз значения временного ряда в требуемый момент времени. Процедура выбора экзогенных факторов и их использование являются составной частью процедуры оценивания параметров модели.

Сложность прогнозирования объёмов грузовых железнодорожных перевозок обусловлена высокой волатильностью и нестационарностью исторических данных об объёмах перевозок, а также наличием в них значительного количества пропусков. К примеру, не для всех пар станций существуют исторические данные, так между некоторыми парами станций перевозки не производятся. В связи с этим, встает задача определения допустимых запросов на прогнозы, в том числе определения горизонта прогнозирования, при котором получаемые прогнозы обладают достаточным уровнем достоверности.

Предлагается набор критериев для оценки качества прогнозов при выборе оптимальной прогностической модели. Качество оценивается с помощью функций потерь, вычисляемых на основе отклонений прогнозов временного ряда от его истинных значений при ретроспективном прогнозе. Дополнительно учитываются оценки достоверности прогноза, максимального горизонта прогнозирования при заданном уровне достоверности, а так же свойства распределения ошибки.

Таблица 1: Вид записи базы данных железнодорожных перевозок

Дата погрузки	Станция отправления	Станция назначения	Кол-во вагонов	Код груза	Род вагона	Суммарный вес груза	Признак маршрутной отправки
2007-01-01	020108	932902	1	1	216	56	9

2. Модели прогнозирования

Основные методы непараметрической регрессии, такие как ядерное сглаживание, сглаживание сплайнами, авторегрессия, скользящее среднее и др., описаны в [4, 5, 6]. Они заключаются в присвоении имеющимся значениям временного ряда весов и комбинации взвешенных значений для получения прогноза. Альтернативный подход к прогнозированию заключается в задании параметрической модели, определяющей значение прогноза при выбранных параметрах. Используются модель авторегрессионного скользящего среднего ARMA, обобщающая ее интегрированная модель авторегрессионного скользящего

среднего ARIMA [7]. Модель ARIMA, широко используемая для прогнозирования нестационарных временных рядов, позволяет с хорошим качеством прогнозировать временные ряды с трендом, а также при небольшой модификации и ряды с сезонной компонентой (модель SARIMA) [7]. В работе [2] была предложена непараметрическая модель гистограммного прогнозирования hist, разработанная для прогнозирования стационарных временных рядов.

Перечисленные модели применяются для прогнозирования одномерных временных рядов, то есть позволяют вычислять прогнозы на основе исключительно исторических данных. Учет дополнительных временных рядов зачастую приводит к повышению качества прогнозов. Модификации перечисленных выше моделей позволяют учесть при прогнозировании дополнительные временные ряды. Для модели ARMA [9, 8], содержащей три аддитивных компоненты: авторегрессионную, скользящее среднее и ошибку, была разработана экзогенная модификация ARMAX [10]. Модель ARMAX включает также комбинацию экзогенных временных рядов в качестве дополнительной аддитивной компоненты. Модель гистограммного прогнозирования hist может быть расширена на случай нескольких переменных путем обобщения методов многомерной квантильной регрессии. Задаче разработки методов многомерной квантильной регрессии посвящен ряд работ [11, 12, 13]. Каждое из предлагаемых решений сфокусировано на сохранении определенных характеристик одномерной квантильной регрессии. В данной работе рассмотрена модель векторной авторегрессии VAR, позволяющая одновременно прогнозировать наборы временных рядов, на каждый из которых накладывается требование стационарности. В случае, когда исследуемые временные ряды нестационарны, необходимо использовать дополнительные инструменты, учитывающие коинтеграцию [14] временных рядов.

Сравнению методов прогнозирования посвящен ряд работ [15, 16, 17, 18], в каждой из которых зафиксированы целевые данные, определяемые решаемой задачей, и масштаб прогнозирования. Например, в работе [15] проводится сравнение моделей прогнозирования данных, содержащих сезонные изменения, на основе помесечных данных об объемах контейнерных грузов в международных портах Тайваня. Рассматриваются модели вида GM(1,1) [19, 20], модель SARIMA, модель тригонометрической регрессии [21], линейная регрессионная модель с фиктивными сезонными переменными и декомпозиция временных рядов на мультипликативные компоненты [22], соответствующие тренду, сезонности и случайным вариациям. Авторы [15] приходят к выводу, что разложение временных рядов на мультипликативные компоненты в среднем является оптимальным подходом для прогнозирования изучаемых временных рядов. В работе [16] сравнивается качество различных моделей сезонности и нелинейности для временных рядов, содержащих данные о поквартальных объемах производства в 18 странах ОЭСР. Точность поточечных прогнозов, полученных в [16] зависит от исследуемых рядов, выбранных горизонтов прогнозирования и сезона. В среднем, согласно [16], линейные модели чаще превосходят нелинейные при небольшом горизонте прогнозирования, в то время как нелинейные модели с более сложным описанием сезонности оптимальны при увеличении горизонта прогнозирования. В работе [17] также проведено сравнение точности линейных и нелинейных моделей при прогнозировании агрегированных объемов продаж.

Основная задача данной работы — описание методики сравнения прогностических алгоритмов при выборе оптимального алгоритма прогнозирования с выбранной детализацией. При этом масштаб прогнозирования варьируется: рассматривается детализация прогнозов по дням, по неделям и по месяцам, а также по парам станций и по парам районов. Каждому из перечисленных масштабов может соответствовать свой оптимальный

алгоритм прогнозирования.

В работе [23] проведено исследование критериев оценки методов прогнозирования на основе экспертных мнений. Точность прогнозов была признана наиболее важным критерием большинством исследователей. Тем не менее, ни один из способов измерения точности не является универсальным, в связи с чем для оценки качества прогнозов, как правило, используются несколько функций ошибки, причем качество сравниваемых моделей зависит от используемой функции ошибки [24]. В данной работе для оценки качества прогнозов используются среднее абсолютное отклонение MAE и среднее относительное абсолютное отклонение MAPE. Предложены несколько дополнительных критериев качества для выбора оптимальной модели, в том числе при прогнозировании временных рядов с переменным масштабом.

Кроме того, полученные в результате исследования методов прогнозирования результаты свидетельствуют о необходимости введения дополнительных критериев качества и возможно, асимметричных функций ошибки при прогнозировании объемов грузовых железнодорожных перевозок. В работе [25] были исследованы свойства прогнозов модели ARIMA при использовании несимметричных функций потерь, и сделан вывод о необходимости разработки специальных методов прогнозирования временных рядов в условии несимметричности функции потерь. Методы квантильной регрессии [26] позволяют находить оптимальный прогноз для несимметричных функций потерь кусочно-линейного вида, но применимы только для стационарных временных рядов и не дают возможности использовать другие виды функций потерь. Алгоритм hist, полученный в результате обобщения методов квантильной регрессии, находит приближенное решение задачи минимизации математического ожидания функции потерь произвольного вида. Для прогнозирования нестационарных временных рядов с асимметричной функцией потерь предлагается комбинация моделей ARIMA+hist: на первом этапе вычисляется несмещенный прогноз с помощью модели ARIMA, затем производится анализ регрессионных остатков модели ARIMA и вычисляется оптимальное смещение прогноза для минимизации ожидаемых потерь с помощью алгоритма hist.

3. Задача прогнозирования

Введем обозначения $\mathbf{x} = [x_1, \dots, x_T]^T$ для прогнозируемого временного ряда, $\mathbf{y}^j = [y_1^j, \dots, y_T^j]^T$ — для экзогенных временных рядов $j = 1, \dots, n$.

Моделью прогнозирования временного ряда \mathbf{x} с учетом экзогенных временных рядов назовем параметрическую функцию $f(\mathbf{x}, \mathbf{Y}, \mathbf{w})$, где $\mathbf{Y} = [\mathbf{y}^1, \dots, \mathbf{y}^j, \dots, \mathbf{y}^n]$. При вычислении прогноза на одну точку вперед требуется найти оценку \hat{x}_{T+1} следующего значения x_{T+1} временного ряда \mathbf{x} , минимизирующую заданную функцию потерь $L(x, \hat{x}) : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$

$$\hat{x}_{T+1} = f(\mathbf{x}, \mathbf{Y}, \hat{\mathbf{w}}).$$

Здесь $\hat{\mathbf{w}}$ — вектор оптимальных параметров модели, минимизирующий суммарные потери

$$(1) \quad \hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{t \in \mathcal{T}} L(x_t, \hat{x}_t)$$

при прогнозе участка временного ряда \mathbf{x} , определяемого набором индексов \mathcal{T} .

Ретроспективное прогнозирование. Процедура ретроспективного прогноза заключается в следующем. Фиксируется длина t_0 временных рядов

$$\mathbf{x}_{1 \div t_0} = [x_1, \dots, x_{t_0}] \text{ и } \mathbf{Y}_{1 \div t_0} = [\mathbf{y}_{1 \div t_0}^1, \dots, \mathbf{y}_{1 \div t_0}^n],$$

на основе которой согласно (1) оптимизируются параметры \mathbf{w} прогностической модели f . С помощью модели с оптимальными параметрами $\hat{\mathbf{w}}$ вычисляются прогнозы

$$\hat{x}_t = f(\mathbf{x}_{1:(t-1)}, \mathbf{Y}_{1:(t-1)}, \hat{\mathbf{w}})$$

в каждой точке $t \in \mathcal{T} = \{t_0 + 1, \dots, T\}$. Для оценки качества полученного прогноза используются функции ошибки MAE (mean average error)

$$(2) \quad \text{MAE} = \frac{\sum_{t \in \mathcal{T}} |x_t - \hat{x}_t|}{\sum_{t \in \mathcal{T}} x_t}$$

и, в случае когда значения x_t прогнозируемого временного ряда \mathbf{x} ненулевые для всех $t \in \mathcal{T} = \{t_0, \dots, T\}$, MAPE (mean average percentage error)

$$(3) \quad \text{MAPE} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{|x_t - \hat{x}_t|}{x_t},$$

вычисленные при ретроспективном прогнозе части ряда, определяемого множеством индексов \mathcal{T} .

4. Описание алгоритмов

В данном разделе описано несколько прогностических моделей: векторной авторегрессии, авторегрессионного скользящего среднего и непараметрическая модель hist. Так как каждая из моделей имеет свои достоинства и недостатки, рассмотрена комбинация моделей ARIMA+hist.

4.1. Модель векторной авторегрессии

В рамках модели VAR(p) векторной авторегрессии порядка p вектор $\mathbf{z}_t = [x_{T+1}, y_t^1, \dots, y_t^n]^\top$ моделируется уравнением

$$\mathbf{z}_t = \boldsymbol{\phi} + \boldsymbol{\Phi}_1 \mathbf{z}_{t-1} + \dots + \boldsymbol{\Phi}_p \mathbf{z}_{t-p} + \boldsymbol{\varepsilon}_t,$$

где $\boldsymbol{\phi} = [\phi_0, \dots, \phi_n] \in \mathbb{R}^{n+1}$ — вектор аддитивных параметров, $\boldsymbol{\Phi}_k = [\phi_{ij}^k] \in \mathbb{R}^{(n+1) \times (n+1)}$ — матрицы параметров модели $i, j = 0, \dots, n$; $\boldsymbol{\varepsilon}_t$ — белый шум с нулевым математическим ожиданием.

В этом случае $\hat{x}_t = f(\mathbf{x}, \mathbf{Y}, \hat{\mathbf{w}})$, где

$$f(\mathbf{x}, \mathbf{Y}, \mathbf{w}) \equiv f(\mathbf{x}_{t-p:t-1}, \mathbf{Y}_{t-p:t-1}, \mathbf{w}) =$$

$$\phi_0 + [\phi_{01}^1, \dots, \phi_{0n}^1] \cdot \mathbf{z}_{t-1} + \dots + [\phi_{01}^p, \dots, \phi_{0n}^p] \cdot \mathbf{z}_{t-p} + \varepsilon_t^0.$$

Таким образом, вектор параметров $\mathbf{w} \in \mathbb{R}^{pn+1}$ имеет вид

$$\mathbf{w} = [\phi_0, \phi_{01}^1, \dots, \phi_{0n}^1, \dots, \phi_{01}^p, \dots, \phi_{0n}^p].$$

4.2. Интегрированная модель авторегрессионного скользящего среднего

В рамках интегрированной модели авторегрессионного скользящего среднего ARIMA(p, d, q) ряд разностей

$$\nabla^d x_t = (1 - L)^d x_t, \quad L^d x_t = x_{t-d}$$

временного ряда \mathbf{x} моделируется авторегрессионным скользящим средним ARMA(p, q)

$$\nabla^d x_t = \alpha + \varphi_1 \nabla^d x_{t-1} + \dots + \varphi_p \nabla^d x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q},$$

где $\alpha = \mu(1 - \varphi_1 - \dots - \varphi_p)$; $\mu, \varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q$ параметры модели, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ — шумовая компонента. Таким образом,

$$\hat{x}_t = f(\mathbf{x}_{t-(p+d):t-1}, \mathbf{w}) = \nabla^d x_t - \sum_{i=1}^d (-1)^i \binom{d}{i} L^i x_t,$$

где

$$\mathbf{w} = [\mu, \varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q]^\top.$$

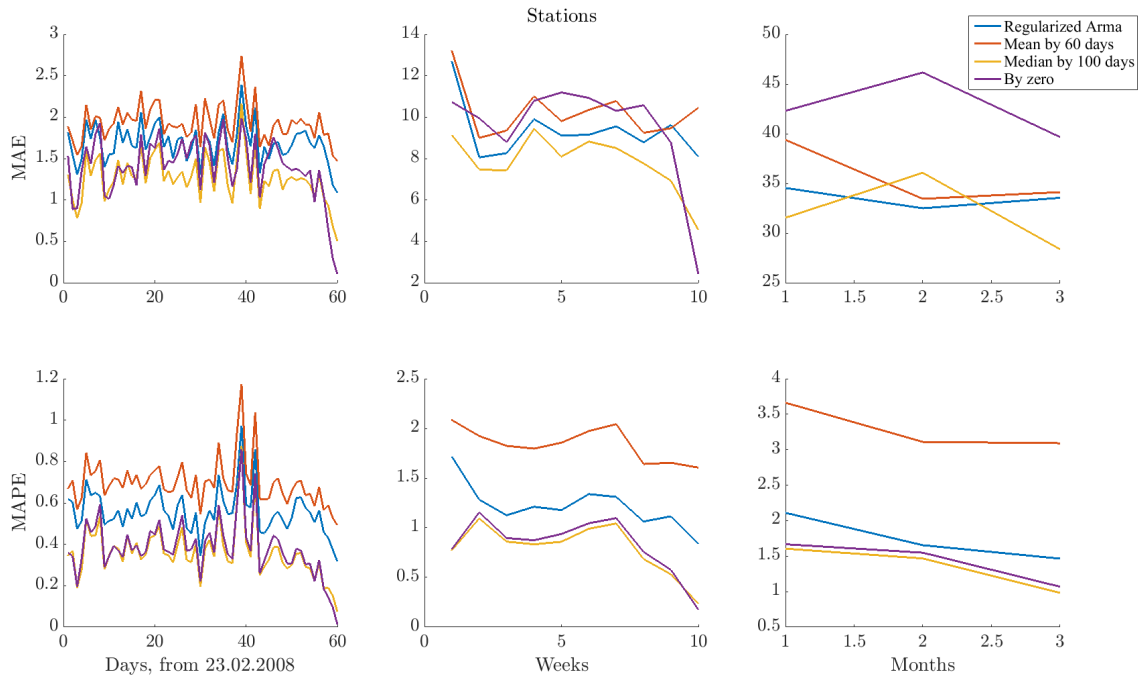


Рис. 1: Ошибка, агрегированная по станциям. Исторические временные ряды.

4.3. Модель гистограммного прогнозирования

Модель hist находит приближенное решение задачи минимизации ожидаемых потерь

$$(4) \quad \hat{x} = \arg \min_{x \in \mathbf{R}^+} \int_0^{+\infty} L(x, u) p(u) du = \arg \min_x \mathbf{E}L(x),$$

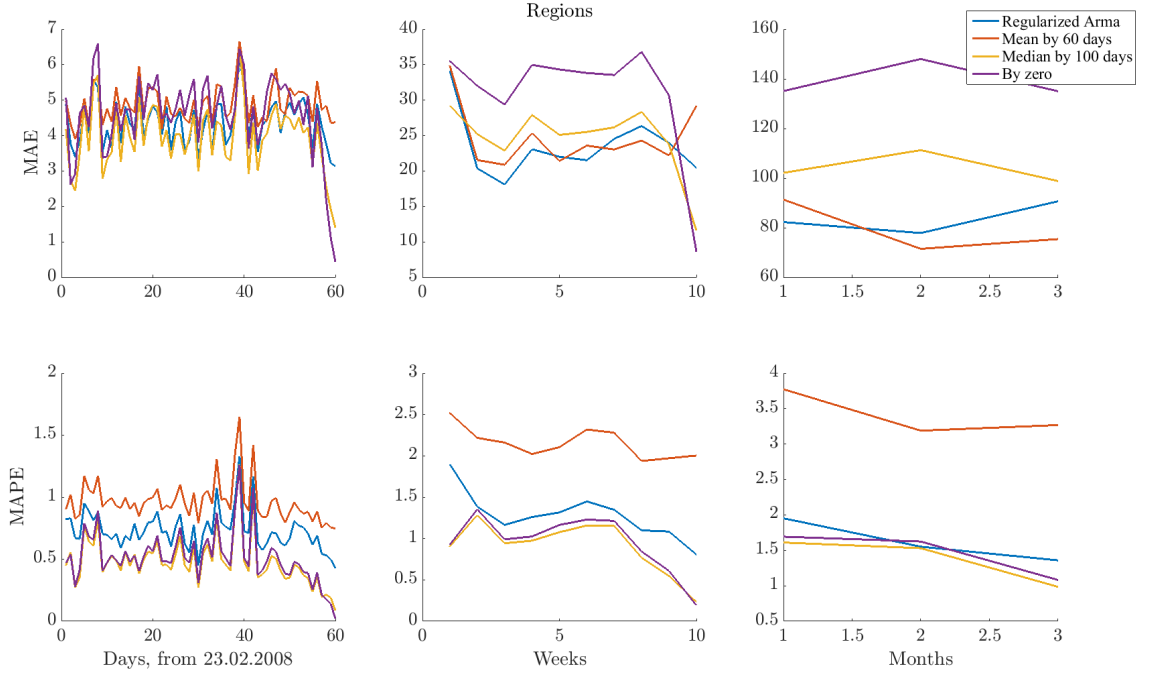


Рис. 2: Ошибка, агрегированная по районам. Исторические временные ряды.

Для этой цели предлагается оценить неизвестную плотность $p(u)$ распределения значений ряда x гистограммой $\hat{p}(u)$ значений ряда. При использовании оценки плотности $\hat{p}(u)$ математическое ожидание потерь $L(x)$ оценивается величиной $L_{\text{hist}}(x)$

$$\mathbf{E}L \approx L_{\text{hist}} = \int_{u_{\min}}^{u_{\max}} L(x, u) \hat{p}(u) du =$$

$$h_1 \int_{u_0}^{u_1} L(x, u) du + \dots + h_N \int_{u_{N-1}}^{u_N} L(x, u) du = \sum_{i=1}^N h_i \int_{u_{i-1}}^{u_i} L(x, u) du,$$

где N — число участков постоянства функции,

$$u_0 = u_{\min} \leq u_1 \leq \dots \leq u_{N-1} \leq u_N = u_{\max}$$

— концы интервалов гистограммы, а h_1, \dots, h_N — значения высоты столбцов гистограммы на этих интервалах. Тогда задача прогнозирования (4) принимает вид

$$(5) \quad \hat{x} = \arg \min_{u_{\min} \leq x \leq u_{\max}} L_{\text{hist}}(x) = \arg \min_{u_{\min} \leq x \leq u_{\max}} \sum_{i=1}^n h_i \int_{u_{i-1}}^{u_i} L(x, u) du.$$

В частных случаях функций потерь $L(x, u)$, например для абсолютного отклонения

$$(6) \quad L(x, u) = |x - u|,$$

использующегося при вычислении ошибок MAE (2) и MAPE (3), или квадратичного отклонения

$$(7) \quad L(x, u) = (x - u)^2$$

существует точное решение задачи минимизации ожидаемых потерь (4). В частности, для абсолютной функции потерь оптимальным является медиана прогнозируемой величины $x(t)$, для квадратичной — ее математическое ожидание. Так как истинное распределение $p(u)$ случайной величины $x(t)$ неизвестно, прогноз алгоритма hist возвращает некоторые оценки медианы и среднего. В описании экспериментов эти оценки приводятся в качестве результатов алгоритма hist.

4.4. Комбинация моделей ARIMA и hist

В случае, когда временной ряд \mathbf{x} не является стационарным, например имеет тренд или сезонную компоненту, нельзя предполагать, что значения временного ряда сгенерированы из одного распределения с плотностью $p(u)$. Прежде чем минимизировать ожидаемые потери в задаче (4), необходимо оценить и исключить из временного ряда все нестационарные особенности. Таким образом прогноз \hat{x} нестационарного временного ряда будет складываться из двух частей: прогноз нестационарной компоненты \hat{x}^{ns} и прогноз стационарной компоненты \hat{x}^{s}

$$\hat{x} = \hat{x}^{\text{ns}} + \hat{x}^{\text{s}}.$$

Прогноз \hat{x}^{ns} , вычисленный с помощью интегрированной модели авторегрессионного скользящего среднего ARIMA(p, d, q) с оптимальными параметрами, учитывает характерные особенности временного ряда $\mathbf{x} = \{x_1, \dots, x_T\}$, но не функции потерь $l(\hat{x}, x_{T+1})$, которая используется в задаче прогнозирования. Для получения оптимального прогноза, необходимо минимизировать ожидаемые потери, как это описано в (4). Для этой цели предлагается использовать регрессионные остатки

$$\mathbf{r} = \{r_1 = x_1 - \hat{x}_1^{\text{ns}}, \dots, r_T = x_T - \hat{x}_T^{\text{ns}}\}$$

интегрированной модели авторегрессионного скользящего среднего ARIMA, которые при выборе оптимальных параметров составляют стационарный временной ряд с некоторой неизвестной плотностью распределения значений $p(u)$. Предлагается построить дополнительный прогноз \hat{x}^{s} для стационарного ряда из регрессионных остатков, который минимизирует математическое ожидание потерь (4) для этого временного ряда. Такая добавка к несмещенному прогнозу \hat{x}^{ns} позволит учесть особенности несимметричной функции потерь.

5. Анализ качества алгоритмов прогнозирования

Сравнение алгоритмов прогнозирования, описанных в разделе 4, проводилось на основе ошибок MAE (2) и MAPE (3). В связи с относительно небольшим размером реальных данных, эксперименты проводились также на синтетических данных, сгенерированных с учетом специфики исследуемых временных рядов.

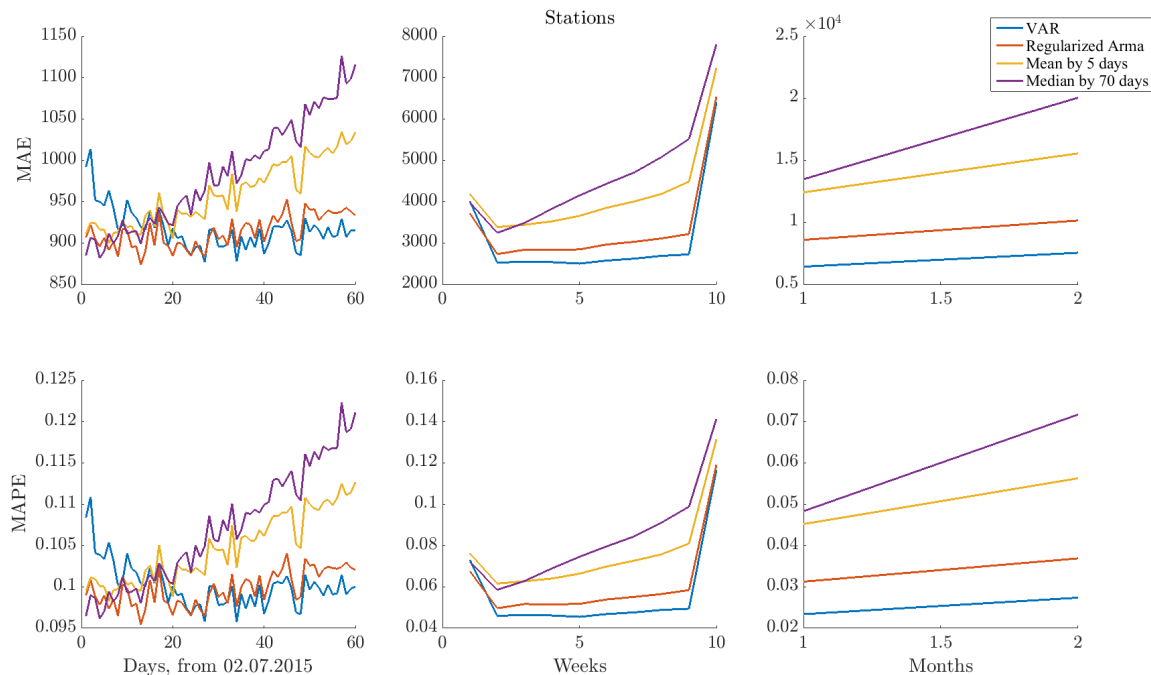


Рис. 3: Ошибка прогнозирования MAE, агрегированная по станциям. Синтетические временные ряды.

5.1. Эксперименты на реальных данных

Экспериментальные данные содержат информацию о железнодорожных перевозках 38 номенклатурных типов грузов. Каждый отсчет $x(t)$ временного ряда \mathbf{x} которых соответствует одному дню и равен суммарному весу (в тоннах) определенного груза, перевезенному между фиксированными пунктами отправления и назначения. Временные ряды были измерены в период с января 2007 года по май 2008 года, в связи с чем временное агрегирование проводилось только по неделям или месяцам: усреднение данных по кварталам или годам сокращает временные ряды до шести или двух точек, соответственно. При запросе на прогноз по неделям или по месяцам выполнялся подневный прогноз, результаты которого затем усреднялись в соответствии с запросом. Согласно требованиям к прогнозу, перечисленным во введении к данной работе, прогноз выполняется для объемов перевозок между фиксированными пунктами отправления и назначения для различных типов грузов. При сравнении алгоритмов прогноз вычислялся для всех пар станций, а также для всех пар районов, имеющих ненулевую историю перевозок.

В таблице 2 приведены значения выбранных функций потерь MAE (2) и MAPE (3), усредненные по типам перевозимых грузов, вычисленные при прогнозировании по парам станций и парам районов с выбранной детализацией по времени. Полужирным шрифтом выделены лучшие результаты при каждом выборе детализации. Параметры прогнозов с помощью среднего и медианы выбраны экспериментально.

Согласно таблице 2, ошибка прогнозирования, как абсолютная, так и относительная, растет при переходе к менее детальным прогнозам. Оптимальным алгоритмом в среднем оказывается прогноз медианой последних ста отсчетов. Также можно отметить, что относительная ошибка MAPE прогнозирования нулями ненамного превосходит ошибку

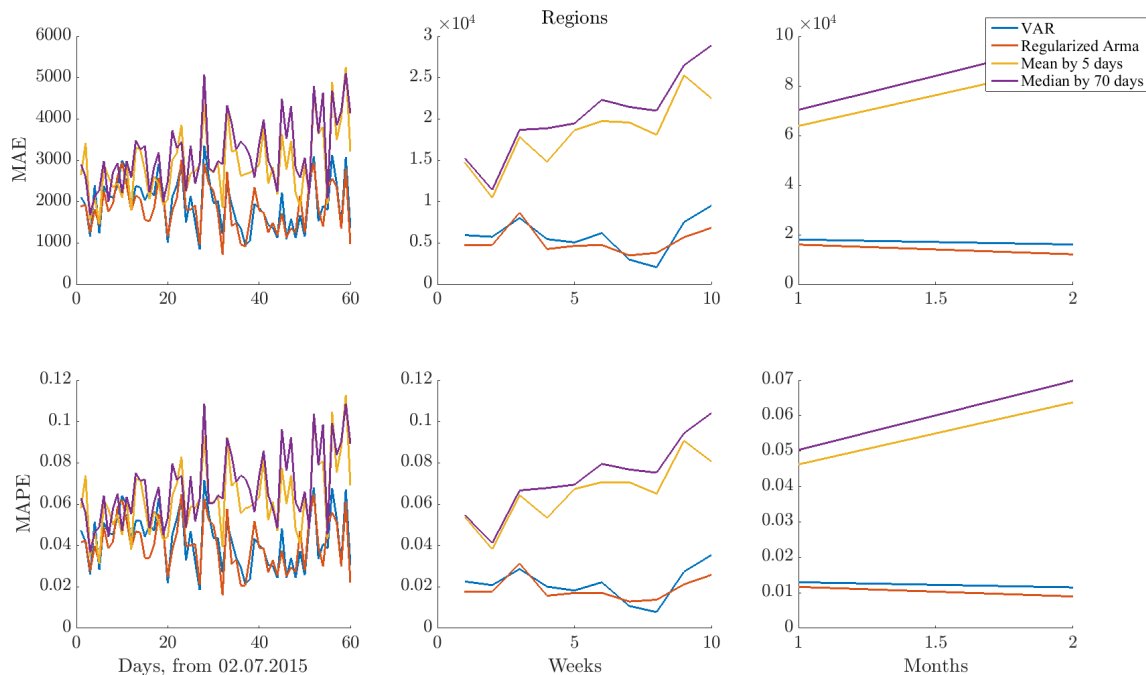


Рис. 4: Ошибка прогнозирования MAPE, агрегированная по районам. Синтетические временные ряды.

при прогнозировании медианой по 100 точкам. Это связано с высокой неравномерностью грузовых перевозок на некоторых направлениях, существенная часть истории некоторых прогнозируемых временных рядов содержит лишь небольшое число ненулевых отсчетов. На синтетических данных, содержащих меньшее количество нулевых значений, такого эффекта не наблюдается. Данный результат свидетельствует о необходимости введения функций потерь, учитывающих специфику решаемой задачи: при использовании стандартных ошибок MAE и MAPE прогноз нулями с математической точки зрения близок к оптимальному, однако на практике такой способ прогнозирования не применим.

На рисунках 1 и 2 изображены графики зависимости ошибок MAE и MAPE прогнозов соответствующих временных рядов от горизонта прогнозирования. В рассмотренных случаях ярко выраженной зависимости от горизонта прогнозирования не наблюдается ни для одного из рассмотренных алгоритмов.

6. Эксперименты на синтетических данных

Для проведения более полных экспериментов был разработан генератор модельных данных, учитывающий особенности структуры грузоперевозок на железнодорожном транспорте, определяемые следующими внешними экспертными данными:

- 1) топологией станций с указанием кодов станций и их принадлежностью к районам;
- 2) информацией о парах станций, используемых для перевозки того или иного товара,
- 3) экзогенными факторами, влияющие на совокупный спрос соответствующего товара;
- 4) историческими данными для экстраполяции структуры перевозок на все области;

Таблица 2: Ошибки прогнозирования, усредненные по типам грузов, исторические данные.

	По станциям			По районам		
	День	Неделя	Месяц	День	Неделя	Месяц
MAE						
ARMA с регуляризацией	1.684	9.323	33.528	4.370	23.456	83.595
Среднее по 60 дням	1.927	10.266	35.650	4.892	24.638	79.417
Медиана по 100 дням	1.276	7.813	31.995	3.974	24.570	104.070
Нулевой прогноз	1.398	9.444	42.710	4.572	30.970	139.383
MAPE						
ARMA с регуляризацией	0.559	1.218	1.743	0.732	1.281	1.621
Среднее по 60 дням	0.698	1.840	3.288	0.963	2.154	3.410
Медиана по 100 дням	0.364	0.790	1.352	0.489	0.904	1.375
Нулевой прогноз	0.382	0.830	1.428	0.518	0.955	1.466

5) другими экспертными требованиями, такими как устойчивость структуры перевозок во времени.

Таблица 3: Ошибки прогнозирования, усредненные по типам грузов, синтетические данные.

	По станциям			По районам		
	День	Неделя	Месяц	День	Неделя	Месяц
MAE						
VAR	917.104	3114.073	6986.463	1946.955	5855.633	17053.388
ARMA с регуляризацией	913.336	3381.220	9366.639	1853.329	5168.458	14056.409
Среднее по 5 дням	959.840	4195.548	13987.420	2892.025	18165.117	76272.813
Медиана по 70 дням	984.188	4623.588	16761.591	3168.165	20385.037	84069.813
MAPE						
VAR	0.100	0.057	0.025	0.042	0.021	0.012
ARMA с регуляризацией	0.100	0.062	0.034	0.040	0.019	0.010
Среднее по 5 дням	0.105	0.076	0.051	0.063	0.065	0.055
Медиана по 70 дням	0.107	0.083	0.060	0.068	0.073	0.060

В таблице 3 приведены значения выбранных функций потерь MAE (2) и MAPE (3), усредненные по типам перевозимых грузов, вычисленные при прогнозировании по парам станций и парам районов с выбранной детализацией по времени детализации по времени. Полужирным шрифтом выделены лучшие результаты при каждом выборе детализации. Как и в случае реальных данных, средняя по типам грузов ошибка MAE растет при переходе к менее детальным прогнозам. Для ошибки MAPE наблюдается обратная зависимость. Оптимальными в этом случае моделями оказались модель векторной авторегрессии при прогнозе по парам станций и авторегрессионного скользящего среднего с регуляризацией параметров модели при прогнозе по парам районов. Оптимальные параметры прогнозирования средним и медианой для синтетических рядов меньше, чем для исторических в связи с наличием более выраженного тренда. На рисунках 3 и 4 изображены графики зависимости ошибок MAE и MAPE прогнозов соответствующих временных рядов от горизонта прогнозирования. При росте горизонта прогнозирования для всех алгоритмов наблюдается увеличение ошибки прогнозирования. Этот эффект также связан с наличием тренда.

7. Дополнительные критерии качества моделей

Для каждого из ретроспективных прогнозов $\hat{x}_t = f(\mathbf{x}_{1:t-1}, \mathbf{Y}_{1:t-1}, \mathbf{w})$, $t \in \mathcal{T}$ вычислим ошибку прогнозирования

$$\varepsilon_t = L(x_t, \hat{x}_t), t \in \mathcal{T}$$

с помощью заданной функции потерь $L(x, \hat{x})$, например линейной (6) или квадратичной (7). Предлагается сравнивать прогностические модели f' , f'' на основе анализа векторов $\boldsymbol{\varepsilon}'$, $\boldsymbol{\varepsilon}'' \in \mathbb{R}^{|\mathcal{T}|}$ ошибок, порождаемых этими моделями, с помощью следующих критериев:

1) Средняя ошибка прогнозирования MAE (MAPE). Модель f' лучше модели f'' , если MAE (MAPE) для f' меньше, чем MAE (MAPE) для f'' . Применение данного критерия описано в разделе “Анализ качества алгоритмов прогнозирования”.

2) Критерий знаков. Модель f' лучше модели f'' , если медиана $\text{med}(\boldsymbol{\varepsilon}')$ вектора $\boldsymbol{\varepsilon}'$, порождаемого f' меньше соответствующей медианы $\text{med}(\boldsymbol{\varepsilon}'')$ для f'' . Достоверность d принятого решения о превосходстве модели f' измеряется с помощью статистики w , равной числу отсчетов t , для которых $\varepsilon'_t < \varepsilon''_t$:

$$d = 1 - \frac{1}{2^{|\mathcal{T}|}} \sum_{t=1}^w \binom{|\mathcal{T}|}{t}, \quad \text{где } w = |t : \varepsilon'_t < \varepsilon''_t|.$$

Оценка достоверности d равна вероятности, что количество отсчетов t , для которых $\varepsilon'_t < \varepsilon''_t$, превосходит w при условии $\text{med}(\boldsymbol{\varepsilon}') \geq \text{med}(\boldsymbol{\varepsilon}'')$. Таким образом, чем больше значение d , тем больше достоверность решения $\text{med}(\boldsymbol{\varepsilon}') < \text{med}(\boldsymbol{\varepsilon}'')$.

3) Критерий согласованности. Прогнозируемые временные ряды имеют иерархическую структуру: каждый прогнозируемый временной ряд может быть либо разбит на несколько временных рядов для получения прогнозов с большей детализацией (нижний уровень иерархии), либо агрегирован по времени или суммирован с другими временными рядами для получения прогноза с меньшей детализацией (верхний уровень иерархии). Критерий согласованности отдает предпочтение той модели, для которой порядок ошибок наименее зависим от детализации: модель f' лучше модели f'' , если разница между ошибкой прогнозирования на верхнем уровне иерархии и суммой ошибок на нижнем уровне для f' меньше, чем для f'' .

8. Заключение

Рассмотрена задача анализа качества прогностических моделей при выборе оптимальной модели для прогнозирования объемов спроса на грузовые железнодорожные перевозки с учетом экзогенных факторов и специфики прикладной области. Проведено сравнение основных параметрических моделей, используемых при прогнозировании временных рядов. Предложена модель непараметрического прогнозирования и способ ее комбинирования с существующей моделью интегрированного авторегрессионного скользящего среднего. Так как результат сравнения моделей зависит от выбора функции ошибки, предлагается расширить процедуру сравнения путем введения дополнительных критериев сравнения. В дополнение к функциям потерь, основанным на точности прогнозов при ретроспективном прогнозировании, предложены дополнительные критерии сравнения моделей, учитывающие статистические требования к прогнозу и специфику прогнозируемых временных рядов, в частности их иерархический характер. В связи с высокой волатильностью прогнозируемых рядов предложен способ оценки горизонта прогнозирования с заданной достоверностью. Таким образом, хотя задача выбора оптимальной модели в рассмотренной

прикладной задаче не имеет однозначного решения, выделен широкий набор подходов к сравнению моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Н.П. Терёшина, В.Г. Галабурда, В.А. Токарев и др.* Экономика железнодорожного транспорта: Учебник для вузов ж.-д. транспорта. М.: УМЦ ЖДТ, 2008.
2. *А. С. Вальков, Е. М. Кожанов, М. М. Медведникова, Ф. И. Хусаинов.* Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным. Машинное обучение и анализ данных, 2012. Vol.1. No. 1. P. 448–465. (*A.S. Valkov, E.M. Kozhanov, M.M. Medvednikova, F.I. Husainov* Non-parametric forecasting of railroad stations occupancy according to historical data. Journal of Machine Learning and Data Analysis, 2012. Vol.1. No. 1. P. 448–465.)
3. *C. Chatfield.* Time Series Forecasting. Chapman & Hall, 2001.
4. *W. Hardle.* Applied Nonparametric Regression. Cambridge University Press, 1992.
5. *B. J. T. Morgan* Applied Stochastic Modelling, Second Edition. Chapman and Hall, 2008.
6. *J. R. Magnus, M. S. Morgan* Methodology and Tacit Knowledge: Two Experiments in Econometrics. Chichester/New York: Wiley, 1990.
7. *Box G., Jenkins G. M., Reinsel G. C.* Time Series Analysis: Forecasting and Control. Prentice-Hall, (third ed.) edition, 1994.
8. *Roopaei M., Zolghadri M., Emadi A.* Economical forecasting by exogenous variables // In IEEE International Conference on Fuzzy Systems, 2008. P. 1491–1495.
9. *P. Whittle.* Prediction and Regulation by Linear Least-Square Methods. University of Minnesota Press, 1983.
10. *Peña D., Sánchez I.* Multifold predictive validation in armax time series models // Journal of the American Statistical Association, 2005. Vol. 100. P. 135–146.
11. *Koltchinskii V.* M-estimation, convexity and quantiles // Annals of Statistics, 1997. Vol. 25. No. 2. P. 435–477.
12. *Serfling R.* Quantile functions for multivariate analysis: Approaches and applications // Statistica Neerlandica, 2002. Vol. 56. P. 214–232.
13. *Kong L., Mizera I.* Quantile tomography: using quantiles with multivariate data // Statistica Sinica, 2012. Vol. 22. P. 1589–1610.
14. *Engle R. F., Granger C. W. J.* Co-integration and error correction: Representation, estimation and testing // Econometrica, 1987. Vol. 55. No. 2. P. 251–276.
15. *Wen-Yi Peng, Ching-Wu Chu.* A comparison of univariate methods for forecasting container throughput volumes // Mathematical and Computer Modelling, 2009. Vol. 50. No. 7-8. P. 1045–1057.

16. *Franses P.H., Van Dijk D.* The forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production // International Journal of Forecasting, 2005. Vol. 21. P. 87–102.
17. *Chu C.W., Zhang G.P.* A comparative study of linear and nonlinear models for aggregate retail sales forecasting // International Journal of Production Economics, 2003. Vol. 86. P. 217–231.
18. *Taylor J.W., De Menezes L.M., McScharry P.E.* A comparison of univariate methods for forecasting electricity demand up to a day ahead // International Journal of Forecasting, 2006. Vol.22. P. 1–16.
19. *Deng J.L.* Introduction grey system theory // Journal of Grey System, 1989. Vol.1. P.1–24.
20. *Tzeng F.M., Yu H.C., Tzeng G.H.* Applied hybrid grey model to forecast seasonal time series // Technological Forecasting and Social Change, 2001. Vol. 67. P. 291–302.
21. *Bowerman B.L. and O'Connell R.T.* Forecasting and Time Series: An Applied Approach. Duxbury Press, Belmont, CA, (3rd ed.) edition, 1993.
22. *M. G. Kendall.* Time-Series. (second ed.) edition, 1976.
23. *Yokum J.T., Armstrong J.S.* Beyond accuracy: Comparison of criteria used to select forecasting methods // International Journal of Forecasting, 1985. Vol. 11. No. 4. P. 591–597.
24. *Makridakis S., Anderson A., Carbone R., Fildes R., Hibon M., Lewandowski R., Newton J., Parzen E., Winkler R.* The accuracy of extrapolation (time series) methods: Results of a forecasting competition // Journal of forecasting, 1982. Vol. 1. P. 111–153.
25. *Patton A. J. and Timmermann A.* Properties of optimal forecasts under asymmetric loss and nonlinearity // Journal of Econometrics, 2007. Vol. 140, No. 2. P. 884–918.
26. *Koenker R.* Quantile regression. Cambridge university press, 2005.
27. *Diebold F. X., Gunther T. A., Tay A. S.* Evaluating Density Forecasts, with Applications to Financial Risk Management // International Economic Review, 1998. Vol. 39. P. 863–883.
28. *Christoffersen P. F., Diebold F. X.* Further results on forecasting and model selection under asymmetric loss // Journal of applied econometrics, 1996. Vol. 11. No. 5. P. 561–571.
29. *Christoffersen P. F., Diebold F. X.* Optimal prediction under asymmetric loss // Econometric theory, 1997. Vol. 13. No. 6. P. 808–817.

Рудаков К.В., ФИЦ ИУ РАН, Вычислительный центр им. А.А. Дородницына, профессор,
Москва, rudakov@ccas.ru

Кузнецов М.П., Московский физико-технический институт, аспирант,
Москва, mikhail.kuznecov@phystech.edu

Мотренко А.П., Московский физико-технический институт, аспирант,
Москва, anastasiya.motrenko@phystech.edu

Стенина М.М., Московский физико-технический институт, студент,
Москва, mmedvednikova@gmail.com

Каширин Д.О., *ФИЦ ИУ РАН, Вычислительный центр им. А.А. Дородницына, инженер-исследователь, Москва, kashirin@forecsys.ru*

Стрижов В.В., *ФИЦ ИУ РАН, Вычислительный центр им. А.А. Дородницына, ведущий научный сотрудник, Москва, strijov@ccas.ru*