Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА и ГОСУДАРСТВЕННОЙ СЛУЖБЫ при Президенте Российской Федерации»

ИНСТИТУТ ЭКОНОМИКИ, МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ
ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ
НАПРАВЛЕНИЕ 38.03.01 ЭКОНОМИКА

Группа ЭО-15-01

Кафедра микроэкономики

Допустить к защите
заведующий кафедрой микроэкономики

_____ М.И. Левин

«____» _____ 201__ г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРОГНОЗИРОВАНИЕ ИЕРАРХИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ

студент-бакалавр
Касьянова Ксения Алексеевна /_____/_____/
(подпись) (дата)

научный руководитель выпускной квалификационной работы
ст. преп. Демешев Борис Борисович /_____/_____/
(подпись) (дата)

МОСКВА
2019 г.

Оглавление

Введение

В анализе данных часто встречаются данные со сложной многоуровневой структурой, точный прогноз которых является одним из ключевых факторов принятия эффективных решений. В связи с этим необходимо использовать уже известные подходы, позволяющие учитывать взаимозависимости прогнозируемых временных рядов, и разрабатывать новые.

С развитием различных социально-экономических процессов, укрепляется и взаимосвязь между ними. Анализ данных с иерархической структурой требуется в микроэкономике (например, при анализе спроса на различные виды товаров в разных городах), макроэкономике (показатели выпуска по регионам по разным отраслям), страховании (анализ рисков попасть аварию, в зависимости от привычек и местонахождения человека), демографии (смертность по регионам и причинам смерти) и т.д. Помимо этого существует и межвременная агрегация временных рядов, часто применяющаяся при прогнозировании.

В данной работе исследуются методы прогнозирования иерархических временных рядов, учитывающие зависимость между уровнями агрегирования и внутри одного уровня. Теоретической основой исследования послужили работы ученых в области анализа данных, прогнозирования и моделирования.

Цель работы: сравнение моделей, учитывающих иерархическую структуру данных, выявление факторов, позволяющих улучшить прогнозы агрегированного временного ряда.

Достижение поставленной цели предполагает постановку и решение следующих задач:

- сбор данных с трехуровневой иерархической структурой;
- выбор моделей для прогнозирования агрегированного ряда;
- сравнение ARIMA моделей с использованием дополнительных регрессоров (ближайшего по метрике корреляции временного ряда третьего уровня, ряда второго уровня или прогноза ряда второго уровня) и без;
- сравнение различных методов комбинирования прогнозов нижних рядов;
- кластеризация временных рядов третьего уровня для получения комбинированных рядов второго уровня (суммирование всех рядов, попавших в один кластер), сравнение прогнозов по ”оригинальным”и ”комбинированным”рядам второго уровня;
- прогнозирование рядов второго и третьего уровня по выбранным моделям, сравнение суммы и оптимальной комбинации этих прогнозов с прогнозом агрегированного временного ряда.

Методы, описанные в данной работе актуальны при необходимости прогнозирования, как агрегированного ряда, так и отдельных компонент, составляющих его, а также получения подтверждения правильности выбора модели для агрегированного ряда. Для анализа были выбраны ряды с определенной структурой, а именно: структура трехуровневая и

иерархическая, причем сам агрегированный ряд и ряды второго уровня можно получить при суммировании рядов третьего уровня.

Практическая значимость работы заключается в том, что при анализе результатов применения изучаемых методов на трех наборах данных (с разной сезонностью, числом наблюдений и рядов на каждом уровне) с использованием перекрестной проверки (кросс-валидации) можно протестировать методы на независимых данных, а следовательно получить более устойчивые выводы.

Данная работа состоит из введения, двух глав основной части, заключения и приложений. В первой главе рассматриваются основные модели прогнозирования иерархических временных рядов. Во второй главе проводится сравнение моделей применительно к собранным данным с требуемой структурой.

1 Модели прогнозирования временных рядов с иерархической структурой

1.1 Обзор литературы

Одним из способов повышения точности прогнозов является агрегирование данных. Один из вариантов - агрегирование временных рядов до составления прогноза, другой - агрегирование самих прогнозов.

С другой стороны информация полученная из аггрегированных рядов может иметь существенное влияние при прогнозировании рядов нижнего уровня, хотя ее использование может сопровождаться некоторыми сложностями.

Для наиболее распространенных моделей прогнозированию существуют альтернативные подходы к анализу временных рядов с иерархической структурой, например, модель векторной авторегрессии (VAR), в которой временные ряды имеют общие параметры или модель байесовской векторной авторегрессии (BVAR), где коэффициенты при различных регрессорах могут иметь общее априорное распределение. В том числе применяются многомерные модели пространства состояний, векторное экспоненциальное сглаживание, а также байесовские подходы, например, их применение к пулу аналогичных временных рядов с помощью ... [Duncan et al. (1993, 2001)]. В таких моделях обычная оценка параметрами объединяется с оцененкой по сгруппированной модели.

Эмпирические результаты показали, что с помощью перечисленных выше методов точность прогноза может быть улучшена, поскольку они используют ковариационную зависимость между временными рядами. Однако использование их связано с выполнением большого числа предпосылок или введения соответствующих ограничений на модель.

Эти методы по крайней мере теоретически могут легко обогнать по качеству прогнозов такие простые подходы, как bottom-up (BU), top-down (TD). Но помимо BU и TD подходов к получению прогнозов аггрегированных рядов, существуют более сложные методы получения оптимальных комбинаций прогонозов, например, ... Однако во многих теоретических и эмпирических работах было замечено, что зачастую более простые методы комбинирования прогнозов оказываются в разы эффективнее, сложных методов, использующих метрики, учитывающие особенности каждого из рядов. Так, например, в статье ... лучший прогноз давало простое взвешивание прогнозов.

1.2 Использование дополнительных регрессоров в ARIMA моделях

Одной из наиболее распространенных моделей прогнозирования взаимозависимых рядов является модель векторной авторегрессии (VAR), однако ее использование может сопро-

вождаться некоторыми сложностями, большое число лагов в модели приводит к громоздким методам вычисления оценок коэффициентов.

В качестве некой альтернативы этому методу можно предложить использование модели ARIMA с дополнительными регрессорами, полученными из прогнозируемого набора данных.

1.3 Взвешенные прогнозы

Тем не менее менее существует мнение, что оптимальным уровнем прозрачности работы центрального банка является некоторый промежуточный уровень. Как это ни парадоксально, высокий уровень прозрачности деятельности ЦБ может привести к неопределенности. Слишком большой объем информации приводит к перегрузке и путанице

1.4 Иерархический Байес

2 Сравнение моделей прогнозирования

2.1 Описание данных

Для анализа необходимо найти наборы данных удовлетворяющие следующим критериям: структура трехуровневая и иерархическая, обладающая свойством аддитивности, т.е. для I рядов второго уровня, каждый из которых делится на J рядов третьего уровня, выполняется:

$$y_t = \sum_{i=1}^{I} y_{i,t} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij,t} \tag{2.1}$$

где $y_{ij,t}, y_{i,t}, y_t$ - значения j-ого ряда третьего уровня, i-ого ряда второго уровня и ряда первого уровня соответственно в момент времени t.

Стоит отметить, что поиск реальных данных, идеально подходящих под такую структуру, затруднен. Обычно для микроэкономических показателей в первую очередь собираются данные по отдельным компонентам, из которых можно получить агрегированные ряды, что удовлетворяет свойству аддитивности, однако получить доступ к таким данным сложно. Альтернативой являются макроэкономические данные, при использовании которых стоит учесть, что в общем случае значение верхнего ряда не будет в точности равно сумме нижних рядов, по причине различий в методологиях сбора рядов разных уровней используемых для избежания двойного учета, неточностей и прочих проблем.

Так например, разбивая ряд ВВП на компоненты по регионам и отраслям, надо учесть, что вообще компоненты будут отражать несколько иной показатель - валовую добавленную стоимостью (ВДС) [1]. Агрегированный ряд, получаемый при суммировании всех ВДС, будет меньше ВВП на величину чистых субсидий на производство и импорт. Такой показатель имеет близкую к единице корреляцию с рядом ВВП, поэтому при точном его прогнозировании мы можем получить представление как об общей динамике всех компонент, составляющих ряд, так и о динамике ряда ВВП. Поскольку целью работы является сравнение моделей, для упрощения будем работать с агрегированными показателями по ВВП, являющиеся простой суммой из рядов нижнего уровня.

Вообще говоря, этот факт учитывается при расчете вклада компонент, составляющих ряд, в процентное изменение агрегированного показателя, не обладающего свойством аддитивности[2]:

[1] Валовая добавленная стоимость определяется как разность между выпуском товаров и услуг и их промежуточным потреблением. ВДС исчисляется на уровне отраслей и отражает образование первичных доходов в результате процесса производства товаров и услуг.

[2] Fox D. R. Concepts and Methods of the U.S. National Income and Product Accounts. Bureau of Economic Analysis (BEA), 2017.

$$C\%\Delta_{i,t} = 100 * \frac{q_{i,t} - q_{i,t-1}}{\sum_j q_{j,t-1}} \tag{2.2}$$

где $q_{i,t}$ - значение i-ого ряда в момент времени t.

Такой показатель позволяет определить изменения в структуре агрегата, что делает его ценным инструментом экономического анализа. Если при прогнозировании с помощью иерархических моделей удастся улучшить прогноз агрегированного ряда, то фактически мы также сможем получить достаточно точные прогнозы показателей вклада каждой компоненты.

Для анализа были выбраны три набора данных с описанными выше свойствами, обладающие разной сезонностью: квартальные, квартальные сезонно сглаженные и месячные данные.

2.1.1 Квартальные данные

Квартальные данные[3] - ряды ВДС по 28 странам Европейского союза (включая Великобританию) в разбивке по основным отраслям[4]:

1) 'A' - сельское хозяйство, лесное хозяйство и рыболовство;
2) 'B' - промышленность (кроме строительства);
3) 'F' - строительство;
4) 'G'- оптовая и розничная торговля, транспорт, услуги общественного питания и т.д.;
5) 'J'- информация и связь;
6) 'K'- финансовая и страховая деятельность;
7) 'L'- операции с недвижимостью;
8) 'M'- профессиональная, научно-техническая, административная деятельность;
9) 'O'- государственное управление, оборона, образование, здравоохранение и социальная работа;
10) 'R'- искусство, развлечения, отдых и другие виды услуг.

Данные собраны за период с 2000-Q1 по 2018-Q3.

Разница между совокупным ВВП всех 28 стран, входящих в состав ЕС и суммой ВДС по всем отраслям для каждого из государств, не превышает 1.5% от ВВП.

[3] Eurostat: European statistics - Database. URL: https://ec.europa.eu/eurostat/data/database.

[4] Eurostat metadata: Annual national accounts (nama10). URL: https://ec.europa.eu/eurostat/cache/metadata/en/nama10_esms.htm.

2.1.2 Квартальные сезонно сглаженные данные

Квартальные сезонно сглаженные данные[5] - это ряды ВДС для каждого из 50 штатов Америки с разбивкой на 21 отрасль. Данные собраны за период с 2005-Q1 по 2018-Q2.

В этом наборе 11 рядов имели пропуски. По четырем из них данные перестали собираться в 2008 году, поэтому эти ряды были исключены целиком. Остальные пропуски были заполнены с помощью экспоненциально взвешенного скользящего среднего с шириной окна 4 [6].

Квартальные оценки ВДС в США пересчитываются с учетом сезонных колебаний следующим образом: BEA оценивает соответствующие коэффициенты сезонной корректировки, после чего удаляет из временного ряда среднее влияние изменений, которые обычно происходят примерно в одно и то же время с одинаковой величиной каждый год. Сезонно несглаженные ряды по этому показателю BEA не публикует.

Показатели по ВДС публикуются в реальном денежном эквиваленте (за базовый год принимается 2012). Надо отметить, что значения реальных показателей ВДС по отраслям не обязательно дают в сумме показатель реального ВДС для каждого штата за интересующий период, поскольку относительные цены, используемые в качестве весов для корректировки показателей по отраслям, отличаются от общего уровня цен используемых для корректировки агрегированного показателя. Для периодов близких к 2012 году, когда значительных отклонений относительных цен от индекса цен по стране не было, показатель ВДС штата совпадает с суммой ВДС по отраслям, хотя вообще эта разница не превышает 0.5% ВВП. Разница между ВВП США и суммой ВДС по отраслям для каждого штата не превышает 2%.

2.1.3 Месячные данные

Месячные данные[7] - показатели рождаемости и смертности по основным причинам в каждом регионе РФ, дающие в сумме естественный прирост населения помесячно. Данные собраны за период с 2006-01 по 2019-01.

Если для каждого из регионов просуммировать по причинам смерти все показатели из набора данных ”Число зарегистрированных умерших по основным классам и отдельным причинам смерти (оперативные данные) значения будут отличаться от показателей набора данных ”Число зарегистрированных умерших (оперативные данные)”. Такое расхождение объясняется тем, что по первому показателю разрабатываются не все причины смерти, а

[5] FRED: Economic Data. URL: https://fred.stlouisfed.org/.

[6] Алгоритм, используемый в пакете R ”imputeTS”имеет адаптивный размер окна: в случае длинных промежутков с пропущенными значениями, размер окна постепенно увеличивается до тех пор, пока не появятся как минимум 2 значения не-NA.

[7] ЕМИСС: государственная статистика: Официальные статистические показатели. URL: https://www.fedstat.ru/.

только основные классы и отдельные причины смерти, имеющие наибольший вес. Также в 2011 году методика разработки показателя была пересмотрена, чтобы соответствовать Международной статистической классификации[8].

В связи с этим для анализа были выявлены три основные группы причин смертности, причем разница между показателем смертности по каждому региону и суммой по всем причинам смертности была добавлена к ряду ”смерть по прочим причинам”:

1) ’УБ’ - смерть из-за болезней (болезней органов дыхания, органов пищеварения, системы кровообращения, инфекционных и паразитарных болезней, новообразований);
2) ’УУ’ - убийство и самоубийство;
3) ’УВ’ - смерть по прочим причинам (отравление алкоголем, транспортные травмы всех видов и внешние причины)

С 2015 года также собираются данные по республике Крым и городу федерального значения Севастополю. Однако данных нужной сезонности и классификации за 2006-2014 годы Держстат Украины не предоставляет, поэтому ряды по этим регионам были исключены из набора данных.

2.2 Описание моделей прогнозирования рядов

Сравнение качества прогнозов

Для сравнения качества прогнозов будут использоваться следующие метрики:

– Средняя ошибка (mean error)

$$ME = \frac{1}{h}\sum_{i=1}^{h}(\hat{y}_{t+i|t} - y_{t+i}) \tag{2.3}$$

– Квадратный корень из среднеквадратичной ошибки (root mean square error)

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(\hat{y}_{t+i|t} - y_{t+i})^2} \tag{2.4}$$

– Средняя абсолютная ошибка в процентах (mean absolute percentage error)

$$MAPE = \frac{1}{h}\sum_{i=1}^{h}\frac{y_{t+i} - \hat{y}_{t+i|t}}{y_{t+i}} * 100\% \tag{2.5}$$

Для выбора параметров модели будем сравнивать точность прогнозов по RMSE. Средняя

[8]Демографический ежегодник России: методические пояснения. URL: http://www.gks.ru/bgd/regl/B17_16/Main.htm.

ошибка позволит понять, насколько хорошо модель улавливает тренд в рядах. MAPE в качестве метрики сравнения точности прогнозов является смешенным показателем, поскольку он будет систематически выбирать модель, прогнозы которой занижены, так как MAPE налагает большие штрафы на отрицательные ошибки, чем на положительные. С расчетом MAPE для набора данных по России возникают трудности, так как в нем имеются нулевые и близкие к нулю значения.

2.2.1 Кросс-валидация

[Evaluating forecast accuracy](https://otexts.com/fpp2/accuracy.html)

2.2.2 Кластеризация

2.2.3 Добавление регрессора

2.3 Сравнение моделей

Функции импульсного отклика валютного курса, индекса РТС и ставки МИАКР на заявление Банка России при такой спецификации и идентифицирующих предположениях будут иметь вид, представленный на рис.

При словесной интервенции Банка России происходит прыжок однодневной ставки МИАКР в течение следующего дня после заявления. В течение четырёх дней ставка МИАКР возвращается к своему прежнему уровню. Вторая спецификация подтверждает выводы, полученные при первой спецификации модели. Импульсные отклики валютного курса и индекса РТС на словесные интервенции Банка России оказываются снова незначимыми.

	ME	RMSE	MAPE	MASE
RW with drift	0.02	34.78	0.82	0.29
Auto ARIMA	70.50	71.55	2.03	0.73
ARIMA	72.67	73.73	2.09	0.75
RW	77.53	100.92	2.52	0.91
ETS	100.50	108.39	2.87	1.04
Theta	103.38	109.50	2.96	1.07
SNaive	134.86	146.04	3.86	1.40

	ME	RMSE	MAPE	MASE
RW with drift	0.02	34.78	0.82	0.29
Auto ARIMA	70.50	71.55	2.03	0.73
ARIMA	72.67	73.73	2.09	0.75
RW	77.53	100.92	2.52	0.91
ETS	100.50	108.39	2.87	1.04
Theta	103.38	109.50	2.96	1.07
SNaive	134.86	146.04	3.86	1.40

Заключение

Взвешивание прогнозов повышает точность прогнозов, по сравнению с невзвешенной суммой прогнозов рядов третьего уровня

Прогнозирование рядов второго уровня дает сопоставимые прогнозы с прогнозом агрегированного ряда, если спецификация модели ряда первого уровня совпадает со спецификацией его компонент [4].

- можно провести анализ на увеличивающихся окнах

- Are there any benefits from using rolling forecasts or recursive filters for prediction?

Примеры иерархических временных рядов:

- Дневные - вклад в какой нибудь индекс - просмотры, сколько заходит на страницу по разделам/по возрасту - Проверить ряды в m3comp

- При улучшении прогноза аггрегированного ряда являющегося суммой нижних рядов, мы сможем сделать вывод, что в среднем прогноз каждого ряда по отдельности стал лучше, что важно при анализе изменения структуры агрегата во времени.

Список литературы

1. Eurostat metadata: Annual national accounts (nama10). — URL: https://ec.europa.eu/eurostat/cache/metadata/en/nama10_esms.htm.
2. Eurostat: European statistics - Database. — URL: https://ec.europa.eu/eurostat/data/database.
3. Fox D. R. Concepts and Methods of the U.S. National Income and Product Accounts. — Bureau of Economic Analysis (BEA), 2017.
4. FRED: Economic Data. — URL: https://fred.stlouisfed.org/.
5. Katz A. J. An Overview of BEA's Source Data and Estimating Methods for Quarterly GDP. — 10th OECD-NBS Workshop on National Accounts, 2006.
6. Moyer B. C., Thompson S. Gross Domestic Product by State Estimation Methodology. — 2017.
7. Демографический ежегодник России: методические пояснения. — URL: http://www.gks.ru/bgd/regl/B17_16/Main.htm.
8. ЕМИСС: государственная статистика: Официальные статистические показатели. — URL: https://www.fedstat.ru/.

Приложение А Программа для поиска и выгрузки статей, касающихся Банка Росии из архива газеты Ведомости (Python)

Остальные месяцы выгружаются аналогичным образом. Теперь необходимо каждую из новостей, имеющих отношение к ЦБ дать оценку. Какой именно характер имеет словесная интервенция, отраженная в данной новости. Если она ведет к ужесточению политики, будем присваивать 1, если к смягчению, то -1. В итоге на выходе будем получать матрицу, каждая строка которой имеет вид [дата, смягчение или ужесточение].

Выпускная квалификационная работа выполнена мной совершенно самостоятельно. Все использованные в работе материалы и концепции из опубликованной научной литературы и других источников имеют ссылки на них.

Объем работы ____ листа(ов).

Объем приложений ____ листа(ов).

« __ » _____ 20 __ г.

_____ / Касьянова Ксения Алексеевна /

(подпись)