

**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	3
1 Модели прогнозирования временных рядов с иерархической структурой	5
1.1 Обзор литературы	5
1.2 Комбинирование прогнозов	6
1.2.1 Метод снизу-вверх	8
1.2.2 Метод сверху-вниз	8
1.2.3 Оптимальная комбинация по OLS	9
2 Сравнение моделей прогнозирования	11
2.1 Описание данных	11
2.1.1 Квартальные данные	12
2.1.2 Квартальные сезонно сплаженные данные	13
2.1.3 Месячные данные	13
2.2 Выбор моделей прогнозирования рядов нижнего уровня	14
2.3 Группировка рядов третьего уровня по регионам, по типу и по метрике расстояния	16
2.4 Сравнение иерархических моделей	17
Заключение	20
Список литературы	21
Приложение А Визуализация временных рядов с трехуровневой структурой	23
Приложение Б Сравнение иерархических моделей	26

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 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.

В результате проведенного анализа были получены три основных вывода:

- эффективность моделей прогнозирования агрегированных рядов с помощью моделей, учитывающих многоуровневую структуру данных, сильно варьируется для разных наборов данных и зависит от структуры рядов-компонент по отдельности;
- комбинирование прогнозов с помощью OLS-корректировки имеет смысл при небольшом числе наблюдений, недостаточном для проведения кросс-валидации, поскольку позволяет устранить сильное отклонение невзвешенной суммы прогнозов от прогноза агрегированного ряда по причине случайного накопления идиосинкритических ошибок;
- предварительная группировка рядов нижнего уровня перед прогнозированием практически во всех случаях приносит положительный результат, по сравнению с прогнозами полученными по трехуровневой модели.

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. В приложении (А) содержатся графики, позволяющие визуализировать структуру данных. В приложении (Б) представлены таблицы, позволяющие сравнить качество прогнозов, полученное по моделям, учитывающим многоуровневую структуру данных с моделью, ее не учитывающей.

1 Модели прогнозирования временных рядов с иерархической структурой

1.1 Обзор литературы

В современной литературе по анализу данных можно выделить несколько основных подходов к прогнозированию временных рядов с иерархической структурой. Основное предположение, на котором построены многие из методов, заключается в том, что при группировании рядов, которые ведут себя одинаково, характерные ошибки внутри групп будут компенсировать друг друга, в то время как более релевантная для прогнозирования индивидуальная динамика будет сохраняться.

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

Информация полученная из агрегированных рядов может иметь существенное влияние при прогнозировании рядов нижнего уровня, хотя ее использование может сопровождаться некоторыми сложностями.

Самыми популярными являются подходы к прогнозированию рядов «сверху-вниз», «снизу-вверх», а также их комбинация. Первый метод заключается в прогнозировании агрегированного ряда с последующим разбиением этого прогноза на компоненты по весам, полученным на основе исторических данных или спрогнозированных пропорций. Второй метод заключается в прогнозировании каждого из дизагрегированных рядов нижнего уровня, а затем использование взвешенной комбинации этих прогнозов для получения прогнозов на более высоких уровнях иерархии.

Принципиально другим методом работы с иерархическими временными рядами является байесовский подход. Для рядов относящихся к одной группе подбирается одинаковое априорное распределение коэффициентов. В простых случаях для эффект априорного распределения такой: если ряды сильно отличаются от сгруппированных, то апостериорное распределение коэффициентов для них становится более узким, и влияние группового априорного распределения уменьшается. Однако, если при изменении одного априорного распределения на другое разумное ответы значительно изменятся, то необходимо собрать больше данных.

Альтернативными моделями, подходящими для прогнозирования временных рядов с многоуровневой структурой являются, например, модель векторной авторегрессии (VAR), в которой временные ряды, относящиеся к одной группе, могут иметь общие параметры, или модель байесовской векторной авторегрессии (BVAR), где коэффициенты в рядах, относящихся к одной группе, могут иметь общее априорное распределение. В том числе применя-

ются многомерные модели пространства состояний, векторное экспоненциальное сглаживание, а также байесовские подходы, например, их применение к пулу аналогичных временных рядов[8].

В таких моделях при оценке параметров нижних рядов, учитывается динамика ряда более высокого уровня. Эмпирические результаты показали, что с помощью перечисленных выше методов точность прогноза может быть улучшена, поскольку при оценке параметров учитывается зависимость между временными рядами, принадлежащими одной группе. Однако использование этих моделей связано с выполнением большого числа предпосылок или введения соответствующих ограничений.

Одной из наиболее распространенных моделей прогнозирования взаимозависимых рядов является модель векторной авторегрессии (VAR), однако ее использование может сопровождаться некоторыми сложностями, например, при большое число лагов в модели приводит значительному росту числа оцениваемых коэффициентов.

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

Однако во многих теоретических и эмпирических работах было замечено, что зачастую более простые методы оказываются в разы эффективнее, сложных моделей с большим числом оцениваемых коэффициентов, учитывающие особенности каждого из рядов. Эти методы, по крайней мере теоретически, по качеству прогнозов могут легко обогнать такие простые подходы, как bottom-up («снизу-вверх»), top-down («сверху-вниз»). Но помимо BU и TD подходов к получению прогнозов агрегированных рядов, существуют более сложные методы получения оптимальных комбинаций прогнозов, например, получение скорректированных прогнозов по OLS модели.

1.2 Комбинирование прогнозов

Данные, к которым этот метод применим, отличаются следующим свойством: когда ряды нижнего уровня группируются, прогнозы каждой группы должны быть равны сумме прогнозов рядов, входящих в группу.

Для любого набора данных, обладающего таким свойством можно построить суммирующую матрицу S , которая отражает структуру иерархии рядов. Примером такой матрицы является:

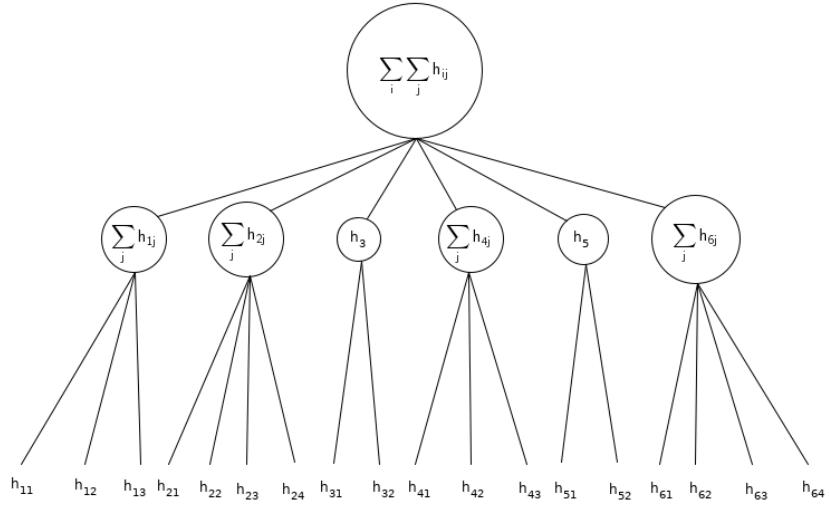


Рисунок 1.1 — Структура временных рядов, необходимых для исследования

$$\begin{bmatrix}
 y_t \\
 y_{1,t} \\
 y_{2,t} \\
 \dots \\
 y_{i,t} \\
 y_{11,t} \\
 y_{12,t} \\
 y_{13,t} \\
 y_{21,t} \\
 y_{22,t} \\
 y_{23,t} \\
 \dots \\
 y_{ij-2,t} \\
 y_{ij-1,t} \\
 y_{ij,t}
 \end{bmatrix} =
 \begin{bmatrix}
 1 & 1 & 1 & 1 & 1 & 1 & \dots & 1 & 1 & 1 \\
 1 & 1 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 1 & 1 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 1 \\
 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1
 \end{bmatrix} \begin{bmatrix}
 y_{11,t} \\
 y_{12,t} \\
 y_{13,t} \\
 y_{21,t} \\
 y_{22,t} \\
 y_{23,t} \\
 \dots \\
 y_{ij-2,t} \\
 y_{ij-1,t} \\
 y_{ij,t}
 \end{bmatrix} \quad (1.1)$$

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 . Схематически такая структура данных представлена на рисунке (1.1).

Переписав уравнение (1.1) в более компактной форме имеем:

$$y_t = Sb_t, \quad (1.2)$$

где y_t - вектор размерности ($n \times 1$) всех наблюдений на всех уровнях иерархии в момент времени t , $n = 1 + i + \sum_i j$, S - суммирующая матрица, отражающая иерархическую структуру данных, b_t - вектор размерности ($m \times 1$) всех наблюдений на самом нижнем уровне иерархии в момент времени t , $m = \sum_i j$

1.2.1 Метод снизу-вверх

Самый распространенный метод получения прогнозов для всех уровней иерархии - bottom-up метод (или метод "снизу-вверх"). Этот заключается в получении независимых прогнозов на h шагов вперед для рядов нижнего уровня иерархии и их агрегирование согласно структуре иерархии S :

$$\tilde{y}_h = S\hat{y}_{K,h}, \quad (1.3)$$

где \tilde{y}_h - собранные с помощью суммирования прогнозы рядов уровней $1...K - 1$ и базовые прогнозы $\hat{y}_{K,h}$.

Основное преимущество этого подхода - при таком методе агрегации не теряется никакая информация. Недостаток метода заключается в том, что ряды нижнего уровня могут быть сильно зашумленными и более сложными для моделирования и прогнозирования[23].

1.2.2 Метод сверху-вниз

Аналогичный предыдущему подход - top-down метод (или метод "сверху-вниз"). Для этого метода необходимо получить базовый прогноз для агрегированного ряда (первого уровня), а также коэффициенты-пропорции на которые агрегированный ряд будет умножаться для получения прогноза нижнего уровня.

"Сверху-вниз" подход может быть эффективнее метода "снизу-вверх" если ряды нижнего уровня сильно зашумлены, но это зависит от того, насколько точными являются прогнозы агрегированного ряда и подобранные пропорции.

Веса-пропорции для этого метода могут, например, подбираться как усредненный за весь период вес i -ого ряда в агрегированном ряде или оцениваться с учетом веса прогноза i -ого ряда в агрегированном.

Также похожим подходом является метод middle-out ("из середины"), объединяющий в себе два предыдущих метода. В этом случае сначала генерируются прогнозы для среднего уровня, а на их основе согласно методам описанным выше получаются прогнозы для верхних и нижних уровней.

1.2.3 Оптимальная комбинация по OLS

Такие популярные методы, как "снизу-вверх" и "сверху-вниз" не учитывают корреляцию между рядами на каждом уровне. Один из методов прогнозирования, позволяющий сделать это, был предложен Хиндманом и др.[18]. Этот метод позволяет получать скорректированные оценки прогнозов агрегированных рядов и их компонент. Основная идея этого метода заключается в получении согласованных на разных уровнях прогнозов на основе независимых между собой прогнозов всех рядов всех уровней.

Подход, описанный Хиндманом[13] предполагает получение прогноза для каждого из рядов иерархи. Предположение следующее - поскольку эти базовые прогнозы генерируются независимо друг от друга, они не могут обладать свойством аддитивности, то есть они не будут суммироваться в соответствии с иерархической структурой, отраженной в матрице S . Данный подход позволяет получить оптимальную комбинацию прогнозов, полученных на основе несогласованных базовых прогнозов, а именно переоценить значения прогнозов для каждого ряда так, чтобы они были максимально приближены к базовым прогнозам, но в то же время удовлетворяли иерархической структуре S .

Идея заключается в представлении базовых прогнозов на h шагов вперед в рамках линейно регрессии:

$$\hat{y}_h = S\beta_h + e_h, \quad (1.4)$$

где \hat{y}_h - вектор базовых прогнозов на h шагов вперед для всех уровней иерархии, S - матрица суммирования, $\beta_h = E[\beta_{n+h}|y_1, \dots, y_n]$ - неизвестное среднее значение будущих значений рядов нижнего уровня, $e_h \sim N(0, \Sigma_h)$ - ошибка регрессии.

В общем случае Σ_h неизвестна. Однако можно показать, что при расчете точечных прогнозов корреляция между рядами не важна.

Если базовые прогнозы приблизительно удовлетворяют иерархической структуре S , то и ошибки также должны приблизительно удовлетворять структуре иерархической агрегации:

$$\hat{e}_h \approx Se_{K,h}, \quad (1.5)$$

где $e_{K,h}$ - вектор ошибок для рядов нижнего уровня.

Такое условие должно выполняться для любого разумного набора прогнозов. При этом предположении, лучшей линейной несмещенной оценкой для β_h является:

$$\hat{\beta}_h = (S'S)^{-1}S'\hat{y}_h. \quad (1.6)$$

Получив оценки $\hat{\beta}_h$ можно пересчитать скорректированные прогнозы для всех уровней иерархии:

$$\tilde{y}_h = S\hat{\beta}_h = S(S'S)^{-1}S'\hat{y}_h. \quad (1.7)$$

Согласно полученной формуле 1.7 пересчитываются скорректированные прогнозы, не зависящие от Σ_h .

Для получения интервальных оценок считается $Var(\tilde{y}_h)$:

$$Var(\tilde{y}_h) = S(S'\Sigma_h^{-1}S)^{-1}S. \quad (1.8)$$

Преимуществом этого метода является то, что этот подход использует доступную в иерархии информацию, учитывает взаимосвязь между прогнозами на всех уровнях иерархии. Этот метод позволяет скорректировать прогнозы на всех уровнях иерархии, и, при условии, что базовые прогнозы являются несмешенными, скорректированные прогнозы также будут несмешенными.

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:

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

где $y_{ij,t}$, $y_{i,t}$, y_t - значения j -го ряда третьего уровня, i -го ряда второго уровня и ряда первого уровня соответственно в момент времени t . Схематически такая структура данных представлена на рисунке (1.1).

Методы иерархического прогнозирования позволяют суммировать прогнозы на каждом уровне, чтобы получить прогнозы на уровне выше. Если данные группируются, прогнозы для агрегированного ряда по группе должны быть равны сумме прогнозов отдельных рядов, входящих в группу.

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 для рядов разных уровней.

Так например, разбивая ряд ВВП на компоненты по регионам и отраслям, надо учесть, что вообще они 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. Так как целью работы является сравнение моделей, для упрощения будем работать с агрегированными показателями по ВВП, являющиеся простой суммой из рядов нижнего уровня.

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.

Вообще говоря, этот факт учитывается при расчете вклада компонент, составляющих ряд, в процентное изменение агрегированного показателя, не обладающего свойством аддитивности [9]:

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

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. В следующих пунктах будут более подробно описаны особенности каждого из наборов данных. Ознакомиться с визуальной презентацией этих наборов можно в приложении (A).

2.1.1 Квартальные данные

Квартальные данные² – ряды ВДС по 28 странам Европейского союза (включая Великобританию) в разбивке по основным отраслям³:

- 'A' - сельское хозяйство, лесное хозяйство и рыболовство;
- 'B' - промышленность (кроме строительства);
- 'F' - строительство;
- 'G' - оптовая и розничная торговля, транспорт, услуги общественного питания и т.д.;
- 'J' - информация и связь;
- 'K' - финансовая и страховая деятельность;
- 'L' - операции с недвижимостью;
- 'M' - профессиональная, научно-техническая, административная деятельность;
- 'O' - государственное управление, оборона, образование, здравоохранение и социальная работа;
- '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.

²Eurostat: European statistics - Database. URL: <https://ec.europa.eu/eurostat/data/database> (дата обращения: 16.02.2019).

³Eurostat metadata: Annual national accounts (nama10). URL: https://ec.europa.eu/eurostat/cache/metadata/en/nama10_esms.htm (дата обращения: 16.02.2019).

2.1.2 Квартальные сезонно сглаженные данные

Квартальные сезонно сглаженные данные⁴ – 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 этих рядов 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 4⁵.

Квартальные оценки ВДС в США пересчитываются с учетом сезонных колебаний следующим образом: ВЕА 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 ВЕА 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 данные⁶ – 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 все показатели из набора данных ”Число зарегистрированных умерших по основным классам и отдельным причинам смерти” просуммировать по причинам смерти, значения будут отличаться от показателей из набора данных ”Число зарегистрированных умерших”. Такое расхождение объясняется тем, что для первового набора разрабатываются ряды только по основным классам и отдельным причинам смерти,

⁴FRED: Economic Data. URL: <https://fred.stlouisfed.org/> (дата обращения: 18.02.2019).

⁵Алгоритм, 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.

⁶ЕМИСС: государственная статистика: Официальные статистические показатели. URL: <https://www.fedstat.ru/> (дата обращения: 10.04.2019).

имеющим наибольший вес. Также в 2011 году методика разработки показателя была пересмотрена, чтобы соответствовать Международной статистической классификации⁷.

Для анализа необходимы ряды, в сумме дающие агрегированный ряд естественного прироста населения. В связи с этим 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". В итоге для каждого региона имеем следующее разбиение:

- 'РО' - число рожденных;
- 'УБ' - число умерших из-за болезней (болезней органов дыхания, органов пищеварения, системы кровообращения, инфекционных и паразитарных болезней, новообразований);
- 'УУ' - число умерших по причине убийства и самоубийства;
- 'УВ' - число умерших 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 по каждому из классов за 2006-2014 годы Держстат Украины не предоставляет, поэтому ряды по этим регионам были исключены из набора данных.

2.2 Выбор моделей прогнозирования рядов нижнего уровня

Для того чтобы определить, можно ли с помощью комбинирования прогнозов получить более точные прогнозы агрегированных рядов необходимо выбрать модель для прогнозирования нижних рядов. Вообще говоря, можно выбирать модели для прогнозирования любого ряда любого уровня независимо друг от друга, оптимизируя, например, метрику качества прогноза. Однако при использовании одной и той же модели, можно увидеть, есть ли зависимость между выбором параметров модели и методом комбинирования рядов или прогнозов.

Эффективность каждого из методов комбинирования будет проверяться на десяти различных моделях: AR с малым числом лагов (с линейным трендом), AR с линейным и с квадратичным трендом, интегрированная AR, ARMA с линейным трендом, ARIMA, ETS с фиксированными параметрами, ARIMA, ETS и TBATS с автоматическим подбором параметров.⁸

Для выбора параметров в моделях применяется кросс-валидация со скользящим ок-

⁷Демографический ежегодник России: методические пояснения. URL: http://www.gks.ru/bgd/regl/B17_16/Main.htm (дата обращения: 09.05.2019).

⁸Автоматический перебор параметров модели осуществляется с помощью функций R 'auto.arima', 'ets' и 'tbats' соответственно.

ном с шагом в одно наблюдение. К рядам нижнего уровня будет применяться модель, для которой среднее по всем подвыборкам RMSE, полученное на кросс-валидации для агрегированного ряда, будет ниже других в классе используемой модели.

Ширина окна для каждого из наборов данных подбиралась с учетом длины ряда и горизонта прогнозирования в два года таким образом, чтобы при проведении перекрестной проверки с шагом в один год получалось не менее пяти подвыборок. В результате

- для квартальных рядов по ВВП ЕС ширина окна 48, прогноз на 8 шагов вперед;
- для сезонно сглаженных рядов по ВВП США ширина окна 28, прогноз на 8 шагов вперед;
- для месячных рядов по естественному приросту РФ ширина окна 84, прогноз на 24 шага вперед.

Различия в ширине окна позволяют сравнить качество прогнозов, полученных по модели на относительно небольшой, средней и большой выборке. Результат перебора параметров для каждой из основных моделей для всех трех наборов данных представлен в таблице (2.1).

Таблица 2.1 — Параметры моделей

	Квартальные	Сезонно сглаженные	Месячные
AR с линейным трендом (с малым числом лагов)	$(p, d, q) = (2, 0, 0)$, $(P, D, Q)_4 = (1, 0, 0)$	$(p, d, q) = (2, 0, 0)$	$(p, d, q) = (2, 0, 0)$, $(P, D, Q)_{12} = (1, 0, 0)$
AR с линейным трендом	$(p, d, q) = (3, 0, 0)$, $(P, D, Q)_4 = (2, 0, 0)$	$(p, d, q) = (4, 0, 0)$	$(p, d, q) = (11, 0, 0)$, $(P, D, Q)_{12} = (2, 0, 0)$
AR с квадратичным трендом	$(p, d, q) = (3, 0, 0)$, $(P, D, Q)_4 = (2, 0, 0)$	$(p, d, q) = (4, 0, 0)$	$(p, d, q) = (11, 0, 0)$, $(P, D, Q)_{12} = (2, 0, 0)$
Интегрированная AR	$(p, d, q) = (3, 1, 0)$, $(P, D, Q)_4 = (2, 1, 0)$	$(p, d, q) = (4, 1, 0)$	$(p, d, q) = (4, 0, 0)$, $(P, D, Q)_{12} = (1, 1, 0)$
ARMA с линейным трендом	$(p, d, q) = (3, 0, 1)$, $(P, D, Q)_4 = (2, 0, 1)$	$(p, d, q) = (4, 0, 1)$	$(p, d, q) = (4, 0, 1)$, $(P, D, Q)_{12} = (1, 0, 1)$
ARIMA	$(p, d, q) = (3, 1, 1)$, $(P, D, Q)_4 = (2, 1, 1)$	$(p, d, q) = (4, 1, 1)$	$(p, d, q) = (4, 1, 1)$, $(P, D, Q)_{12} = (1, 1, 1)$
ARIMA с автоматическим подбором параметров	$\lambda = 1$	$\lambda = 1$	$\lambda = 1$
ETS с фиксированными параметрами	$(E, T, S) = (M, M, M)$ $\lambda = 1$	$(E, T, S) = (A, A, A)$ $\lambda = 1$	$(E, T, S) = (A, Ad, A)$ $\lambda = 1$
ETS с автоматическим подбором параметров	$\lambda = 1$	$\lambda = 1$	$\lambda = 1$
TBATS	$\lambda = 1, T = A$	$\lambda = 1, T = A$	$\lambda = 1, T = Ad$

Для сравнения качества прогнозов будут использоваться следующие метрики:

– средняя ошибка (mean error):

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

– квадратный корень из среднеквадратичной ошибки (root mean square error):

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

– средняя абсолютная ошибка в процентах (mean absolute percentage error):

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

Основной метрикой сравнения точности прогнозов моделей будет RMSE. Средняя ошибка (ME) позволит понять, насколько хорошо модель улавливает тренд в рядах. MAPE в качестве метрики сравнения точности прогнозов является смещенным показателем, поскольку он будет систематически выбирать модель, прогнозы которой занижены, так как на отрицательные ошибки налагаются большие штрафы, чем на положительные. Но зато MAPE позволяет сравнивать улучшение качества прогнозов для разных наборов данных, хотя для набора данных по России MAPE не является показательной метрикой, так как в нем имеются нулевые и близкие к нулю значения.

2.3 Группировка рядов третьего уровня по регионам, по типу и по метрике расстояния

Для данного исследования подбирались наборы данных с трехуровневой структурой, обладающие свойством аддитивности. Это позволяет проверить, можно ли улучшить прогноз агрегированного ряда используя комбинации прогнозов каждого из рядов третьего уровня или такое разбиение рядов на слишком большое число компонент излишне, поскольку приводит к тому, что индивидуальные ошибки прогнозов каждого ряда могут накапливаться, что ведет к ухудшению прогноза агрегированного ряда. Если предположить, что разбиение агрегата на компоненты действительно позволяет учесть неоднородность составляющих агрегированного ряда, но оценка большого числа рядов неизбежно приводит к тому, что идиосинкразические ошибки в сумме растут, то необходимо найти компромисс между двумя этими эффектами.

Очевидно, что можно сгруппировать ряды по территориальному признаку (по странам, штатам или регионам) или по типам (для ВВП группы по отраслям и для естественного прироста отдельно ряды по рождаемости, отдельно по основным причинам смерти). Адди-

тивность позволяет получить ряды второго уровня просто просуммировав ряды, входящие в одну группу.

Альтернативным способом является получение групп с помощью кластеризация нормированных рядов по метрике евклидова расстояния. Для этого использовался алгоритм иерархической кластеризации, реализованный в пакете 'dtwclust'. Оптимальное число кластеров выбиралось так, чтобы максимизировать значение метрики силуэта. Для всех трех наборов данных оптимальное число кластеров - 25. Визуализацию рядов попавших в один кластер можно увидеть на рисунке (A.3).

Надо учесть, что при использовании процедуры перекрестной проверки необходимо было на каждой итерации получать свою группировку на кластеры, но для экономии времени проведем кластеризацию на рядах полной длины.

2.4 Сравнение иерархических моделей

Для сравнения моделей используется следующая процедура:

- разбиение рядов на подвыборки для выполнения перекрестной проверки со скользящим окном с шагом в один год: для квартальных данных число подвыборок равно 6, для сезонно сглаженных - 6, для месячных - 5;
- модели описанные в таблице (2.1) используются для получения прогноза на 2 года вперед для каждого ряда, каждого уровня (отдельно с помощью пакета 'hts' оценивается трехуровневая модель, отдельно три двухуровневые: сгруппированные по регионам, по классам или по кластерам);
- на каждой итерации перекрестной проверки считается RMSFE для агрегированного ряда, RMSFE для невзвешенной суммы всех прогнозов нижнего ряда и RMSFE для скорректированной по OLS суммы всех прогнозов;
- для каждого набора данных RMSFE усредняется по всем подвыборкам и считается процентное изменение RMSFE для невзвешенной суммы всех прогнозов нижнего ряда и RMSFE для скорректированной по OLS суммы всех прогнозов по сравнению с RMSFE для агрегированного ряда;
- полученные значения для трехуровневой и двухуровневых моделей сортируются по RMSFE для агрегированного ряда.

По результатам полученным с помощью описанной процедуры формируются таблицы на которых можно увидеть, как изменился прогноз агрегированного ряда при использовании моделей, учитывающих многоуровневую структуру данных (Приложение (Б)). Для наглядности при относительном уменьшении RMSFE ячейка таблицы окрашивается в зеленый, при относительном увеличении - в красный. Цвета также отличаются по интенсивности, чем

ярче, тем больше отклонение от RMSFE полученного по прогнозам для агрегированного ряда.

При анализе полученных показателей и визуального представления наборов данных (Приложение (A)) были получены три основных вывода:

- эффективность моделей прогнозирования агрегированных рядов с помощью моделей, учитывающих многоуровневую структуру данных, сильно варьируется для разных наборов данных и зависит от структуры рядов-компонент по отдельности;
- комбинирование прогнозов с помощью OLS-корректировки имеет смысл при небольшом числе наблюдений, недостаточном для проведения кросс-валидации, поскольку позволяет устранить сильное отклонение невзвешенной суммы прогнозов от прогноза агрегированного ряда по причине случайного накопления идиосинкразических ошибок;
- предварительная группировка рядов нижнего уровня перед прогнозированием практически во всех случаях приносит положительный результат, по сравнению с прогнозами полученными по трехуровневой модели.

Для невзвешенных прогнозов результаты неоднозначны: для квартальных данных в большинстве случаев при использовании иерархических моделей наблюдается ухудшение по сравнению прогнозом агрегированного ряда, для сезонно сглаженных рядов для большинства моделей наблюдается улучшение прогнозов, а для месячных на некоторых моделях наблюдается улучшение для всех вариантов структуры (трех- и двухуровневой), на некоторых - ухудшение.

Возможно это объясняется следующими фактами. Для квартальных данных все ряды имеют примерно одинаковую структуру и ошибки прогнозов не уравновешивают друг друга, а накапливаются. Для сезонно сглаженных рядов, если посмотреть на метрику ME, можно заметить, что на всех итерациях перекрестной проверки модель занижает прогнозы, но при прогнозировании отдельных компонент можно с этим бороться, поскольку только малая часть прогнозов рядов приводит к тому что тренд недооценивается, и эти прогнозы выравниваются большим числом прогнозов улавливающих положительный тренд. Для месячных данных причина неоднозначных результатов заключается в том, что примерно четверть рядов имеет V-образный тренд (ряды по рождаемости), половина положительный линейный тренд, четверть отрицательный линейный тренд (что видно на рисунке (A.2), где ряды группируются по типу). Соответственно для некоторых моделей хорошие прогнозы имела большая по размеру группа, а для некоторых меньшая.

Прогнозы полученные с помощью OLS корректировки в для квартальных и сезонно сглаженных рядов в большинстве случаев не отличаются от прогнозов полученных при прогнозировании агрегированного ряда. Доля прогнозов имеющих распределение отлича-

ющееся от распределения прогнозов верхних рядов оказалась небольшой, поэтому оценки для агрегата скорректировались незначительно.

Для этих наборов данных для невзвешенных прогнозов наблюдалось резкое ухудшение для квартальных и резкое улучшение для сезонно сглаженных рядов, но при корректировке эти резкие изменения сгладились. Учитывая это можно сказать, что при небольшом числе наблюдений, недостаточном для проведения перекрестной проверки модели, стоит использовать OLS корректировку прогнозов, которая позволит избежать резких отклонений от настоящих значений, произошедших из-за случайного накопления ошибок, которое может привести, как к положительному, так и к отрицательному результату.

Стоит заметить, что для месячных данных при OLS корректировке в отличие от невзвешенной суммы прогнозов наблюдается улучшение прогноза агрегированного ряда практически для любой модели. Исключение составляет только модель с квадратичным трендом. Причина этому заключается в том, что только четверть рядов имеют V-образную форму, такую же, как у агрегированного ряда. Поэтому корректировка прогнозов агрегированного ряда на априори худшие прогнозы, имеющие сильно отличающееся распределение приводит к тому, что прогнозы сильно корректируются в сторону ухудшения.

По той же причине для AR модели с большим числом лагов с линейным трендом для модели с трехуровневой структурой наблюдается ухудшение, поскольку треть рядов прогнозируется по подходящей модели, а четверть по распределению совпадает с некачественным агрегированным прогнозом.

Если сравнивать эффективность использования различных видов группировки можно сказать, что в среднем ее использование перед получением прогнозов отдельных рядов приносит положительный результат. Причем для квартальных данных лучше всего работает группировка по отраслям (как со взвешиванием, так и без), для сезонно сглаженных для обоих способов лучше заметное улучшение наблюдается при группировке по кластерам, а для месячных данных для невзвешенных прогнозов более эффективна группировка по кластерам, а для взвешенных - группировка по типам.

Во всех случаях группировка по территориальному признаку оказывалась чуть хуже других типов группировок, но по отношению к трехуровневой модели без группировки улучшение происходило не во всех случаях.

Вообще говоря, OLS корректировка происходила с учетом второго уровня собираемого именно по территориальному признаку, поэтому имело бы смысл проводить OLS корректировку по рядам третьего уровня с более подходящей для каждого из наборов данных группировкой второго уровня.

Возможная причина таких результатов заключается в том, что при группировке по территориальному признаку ряды приобретают некую обособленность друг от друга, которая в половине случаев позволяет улучшить прогнозы по сравнению с агрегированным

рядов, поскольку используется дополнительная информация о ряде, а в половине добавляет непрогнозируемую ошибку, которая накапливается при суммировании.

При анализе рисунка (A.2) группировка по типу приводит к тому, что в большинстве случаев ряды с общей тенденцией попадают в относительно небольшое число классов, причем ряды похожи друг на друга не только статистически, но и экономический смысл у них один, а соответственно и реакция их на внешние непрогнозируемые шоки с большей вероятностью будет похожей.

Группировка по кластерам в общем-то приносит положительные результаты, однако нельзя забывать, что на определенный момент времени ряды могли случайно попасть в один кластер. Поскольку согласно выбранной метрике кластеризации учитывалась только близость нормированных рядов без учета каких-либо экономических факторов. По этой причине то что в некоторых случаях прогнозы полученные после группировки по кластерам оказывалась хуже других, можно объяснить тем, что алгоритм уловил зависимости, которых на самом деле нет.

Заключение

В этом исследовании продемонстрированы различные подходы к прогнозированию временных рядов с многоуровневой иерархической структурой, позволяющие учитывать взаимозависимости как между самими рядами, так и между прогнозами. Эффективность этих методов проверялась на трех различных наборах данных, длина рядов в которых позволяла проводить перекрестную проверку.

В результате проведенного анализа были получены три основных вывода: эффективность моделей прогнозирования агрегированных рядов с помощью моделей, учитывающих многоуровневую структуру данных, сильно варьируется для разных наборов данных и зависит от структуры рядов-компонент по отдельности; комбинирование прогнозов с помощью OLS-корректировки позволяет устраниТЬ случайное отклонение невзвешенной суммы прогнозов от прогноза агрегированного ряда из-за накопления идиосинкразических ошибок; предварительная группировка с целью снижения вероятности накопления ошибок рядов нижнего уровня приносит положительный результат.

Отметим, что результат применения какого-либо метода сильно зависит от характеристик данных, их свойств и структуры. Целью этого исследования было улучшить прогнозы агрегированного ряда. Однако проведенное исследование показало, что использование иерархических моделей может помочь при анализе сложной структуры специфических наборов данных.

Список литературы

1. Aghabozorgi S., Shirkhorshidi A. S., Wah T. Y. Time-series clustering—A decade review // *Information Systems*. — 2015. — Vol. 53. — P. 16–38.
2. Astakhova N. N., Demidova L. A., Nikulchev E. V. Forecasting method for grouped time series with the use of k-means algorithm // arXiv preprint arXiv:1509.04705. — 2015.
3. Athanasopoulos G., Ahmed R. A., Hyndman R. J. Hierarchical forecasts for Australian domestic tourism // *International Journal of Forecasting*. — 2009. — Vol. 25, № 1. — P. 146–166.
4. Clark J. S. Uncertainty and variability in demography and population growth: a hierarchical approach // *Ecology*. — 2003. — Vol. 84, № 6. — P. 1370–1381.
5. Cobb M. Forecasting Economic Aggregates Using Dynamic Component Grouping // University Library of Munich. — 2017. — Vol. 8. — P. 3–39.
6. Cobb M. Joint forecast combination of macroeconomic aggregates and their components // University Library of Munich. — 2017. — Vol. 3. — P. 2–44.
7. Diebold F. X., Pauly P. The use of prior information in forecast combination // *International Journal of Forecasting*. — 1990. — Vol. 6, № 4. — P. 503–508.
8. Duncan G. T., Gorr W. L., Szczypula J. Forecasting analogous time series // *Principles of forecasting*. — 2001. — Vol. 3. — P. 195–213.
9. Fox D. R. Concepts and Methods of the U.S. National Income and Product Accounts. — Bureau of Economic Analysis (BEA), 2017. — P. 27–39.
10. Gelman A. Multilevel (hierarchical) modeling: what it can and cannot do // *Technometrics*. — 2006. — Vol. 48, № 3. — P. 432–435.
11. Green K. C., Armstrong J. S. Structured analogies for forecasting // *International Journal of Forecasting*. — 2007. — Vol. 23, № 3. — P. 365–376.
12. Hyndman R. J., Athanasopoulos G. *Forecasting: principles and practice*. — OTexts, 2018. — P. 456–523.
13. Hyndman R. J., Lee A. J., Wang E. Fast computation of reconciled forecasts for hierarchical and grouped time series // *Computational Statistics & Data Analysis*. — 2016. — Vol. 97. — P. 16–32.
14. Katz A. J. An Overview of BEA's Source Data and Estimating Methods for Quarterly GDP. — 10th OECD-NBS Workshop on National Accounts, 2006. — P. 145–189.
15. Makridakis S., Hibon M. The M3-Competition: results, conclusions and implications // *International journal of forecasting*. — 2000. — Vol. 16, № 4. — P. 451–476.
16. McNeish D., Wentzel K. R. Accommodating small sample sizes in three-level models when the third level is incidental // *Multivariate behavioral research*. — 2017. — Vol. 52, № 2. — P. 200–215.
17. Moyer B. C., Thompson S. Gross Domestic Product by State Estimation Methodology. — 2017. — P. 34–49.
18. Optimal combination forecasts for hierarchical time series / R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, H. L. Shang // *Computational Statistics & Data Analysis*. — 2011. — Vol. 55, № 9. — P. 2579–2589.

19. Shang H. L., Hyndman R. J. Grouped functional time series forecasting: an application to age-specific mortality rates // *Journal of Computational and Graphical Statistics*. — 2017. — Vol. 26, № 2. — P. 330–343.
20. Shang H. L., Smith P. W. Grouped time-series forecasting with an application to regional infant mortality counts // *Social Statistics & Demography*. — 2013. — Vol. 40. — P. 5–20.
21. Stegmueller D. How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches // *American Journal of Political Science*. — 2013. — Vol. 57, № 3. — P. 748–761.
22. Tobias J. L. Forecasting output growth rates and median output growth rates: A hierarchical Bayesian approach // *Journal of Forecasting*. — 2001. — Vol. 20, № 5. — P. 297–314.
23. Weiss C. Essays in Hierarchical Time Series Forecasting and Forecast Combination. — 2018.
24. Zellner A., Hong C. Forecasting international growth rates using Bayesian shrinkage and other procedures // *Journal of Econometrics*. — 1989. — Vol. 40, № 1. — P. 183–202.

Приложение А Визуализация временных рядов с трехуровневой структурой

Рисунок А.1 — Временные ряды полученные путем агрегирования рядов третьего уровня

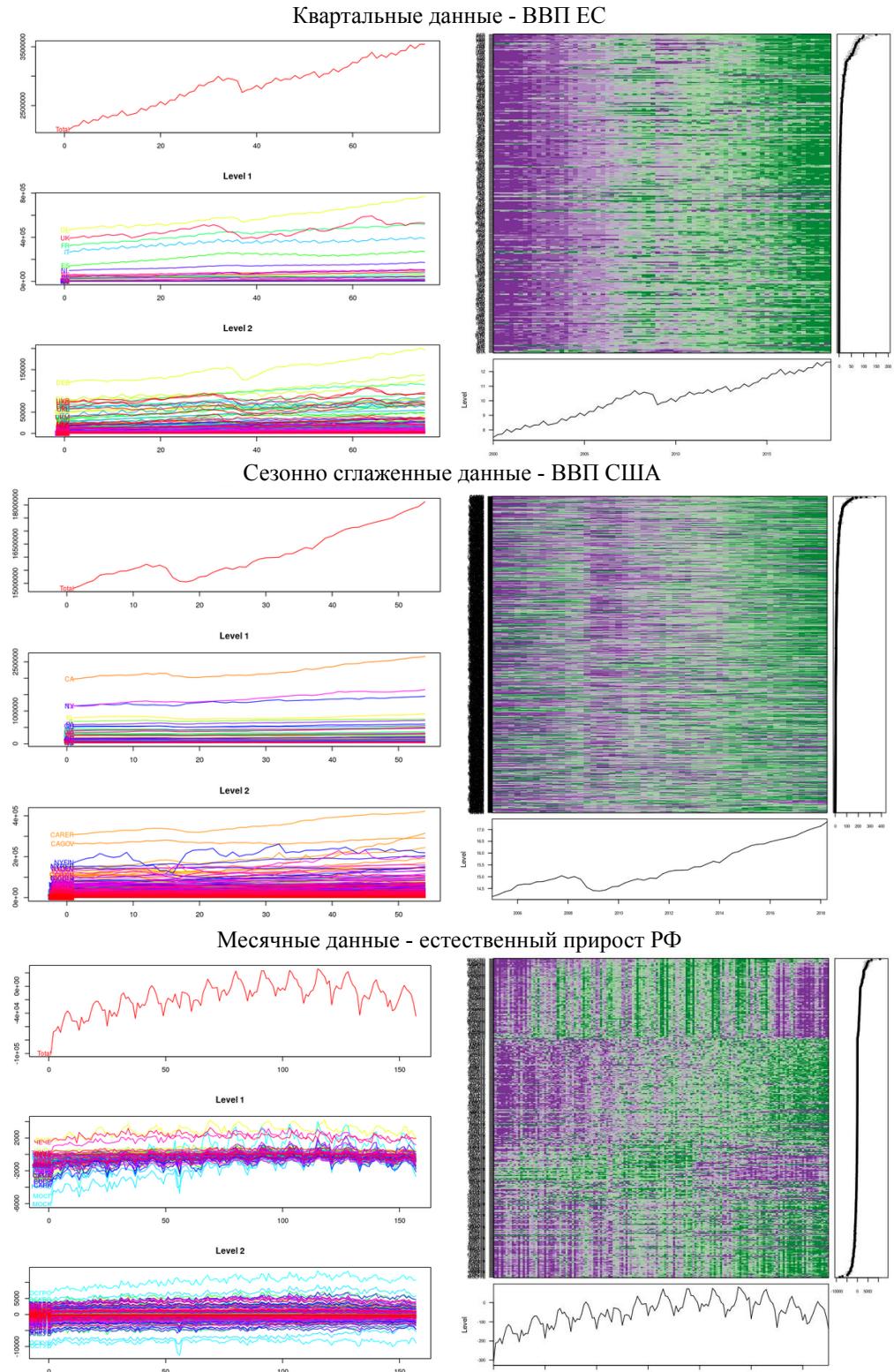


Рисунок А.2 — Временные ряды сгруппированные по территориальному признаку и по типу

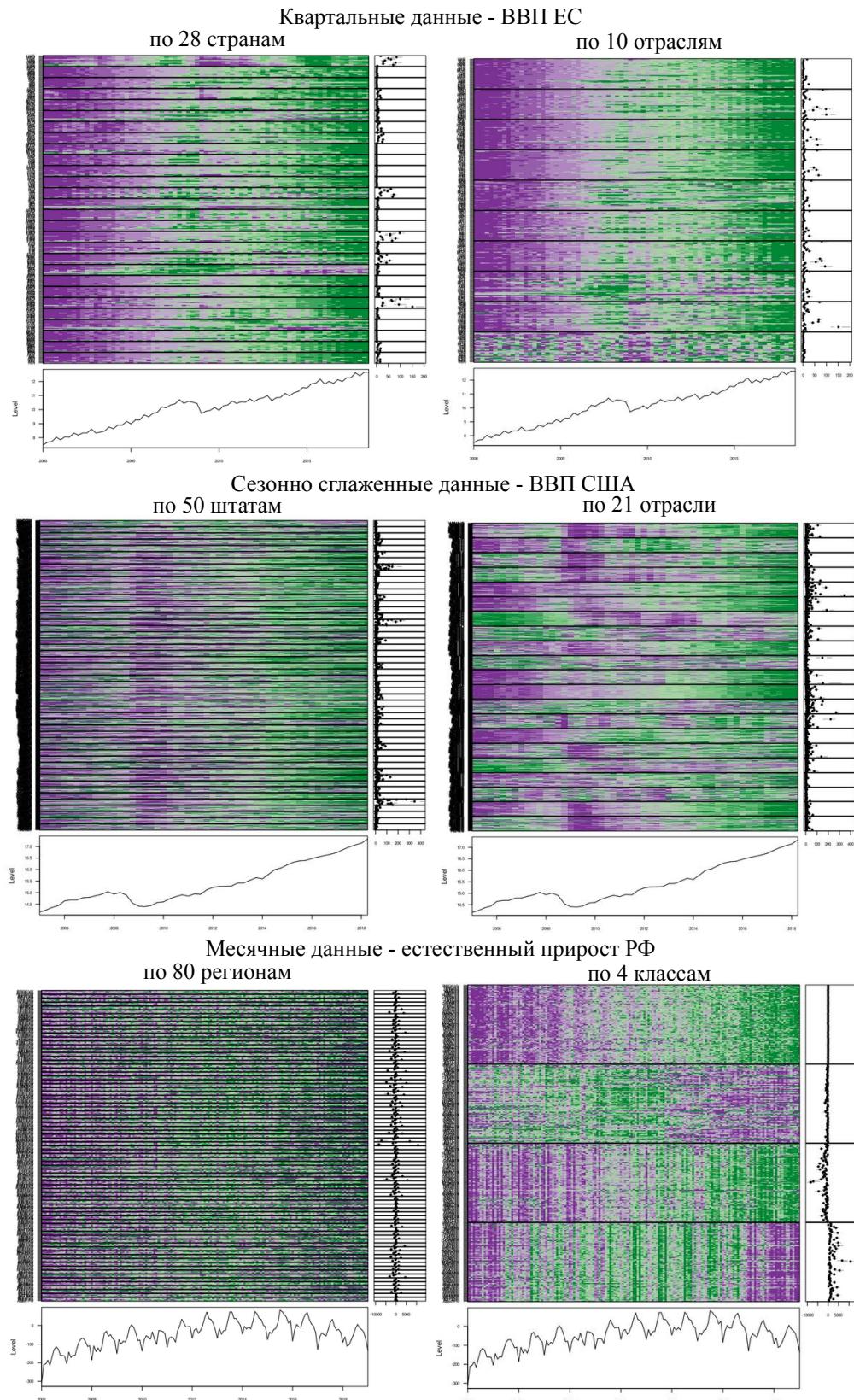
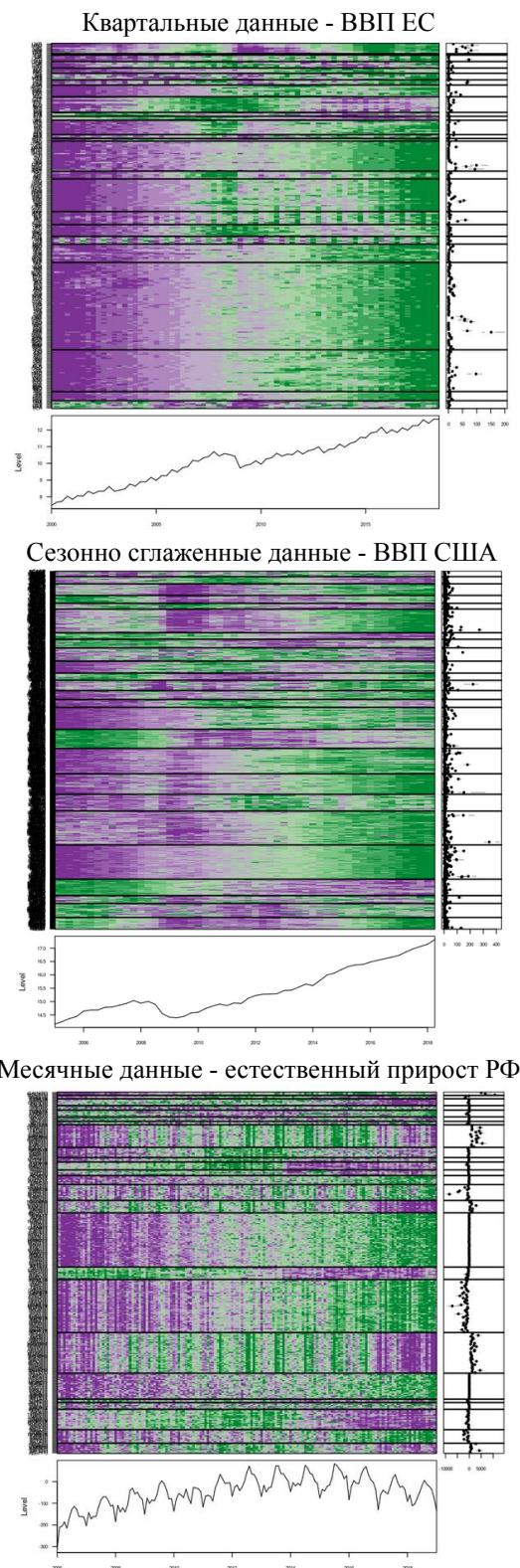
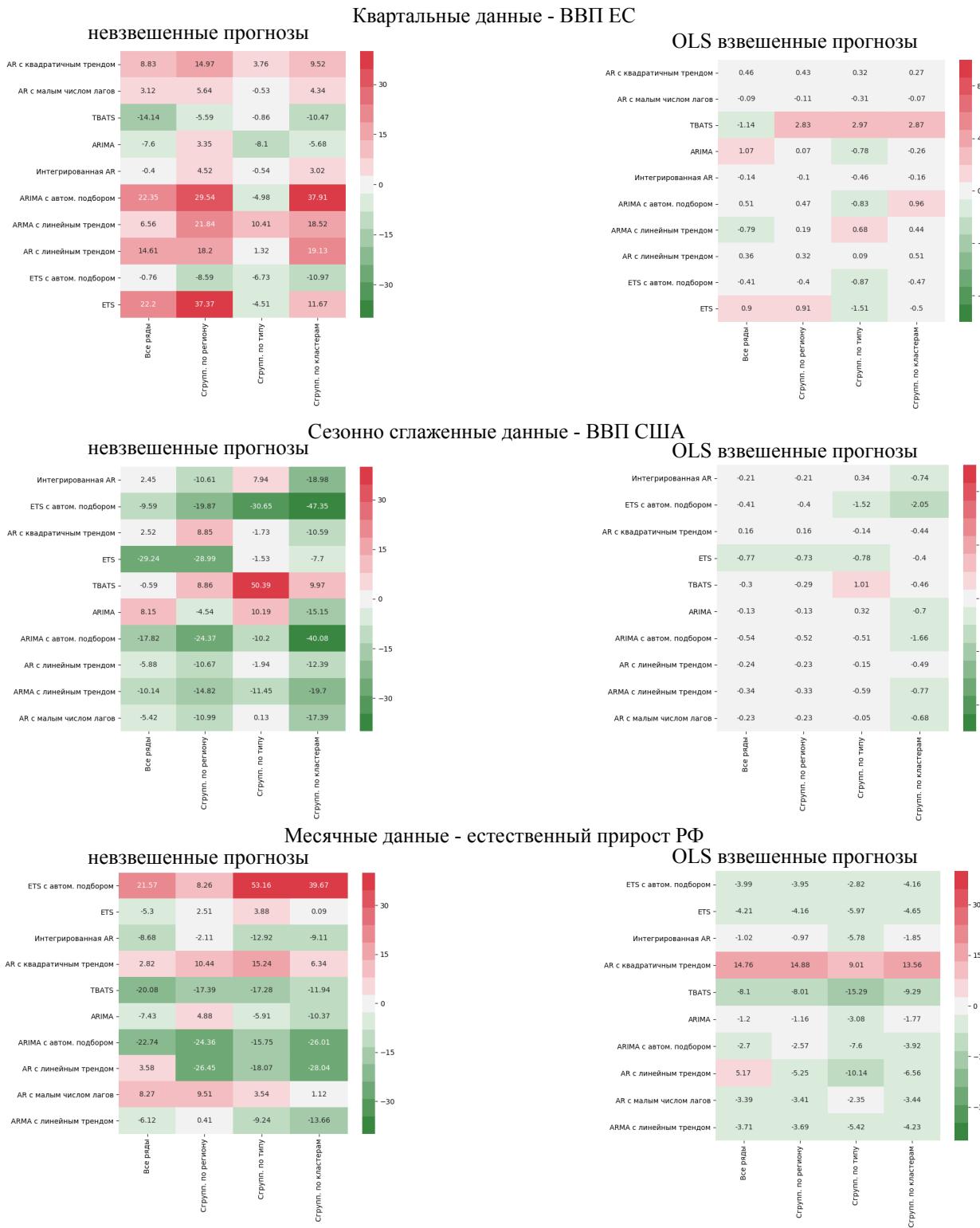


Рисунок А.3 — Временные ряды сгруппированные по метрике расстояния



Приложение Б Сравнение иерархических моделей

Рисунок Б.1 — Таблицы, указывающие на процентное изменение RMSE иерархических моделей по сравнению с RMSE моделей прогнозирования агрегированного ряда



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.

(подпись) / Касьянова Ксения Алексеевна /