

**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 Взвешенные прогнозы	6
2 Сравнение моделей прогнозирования	9
2.1 Описание данных	9
2.1.1 Квартальные данные	10
2.1.2 Квартальные сезонно сплаженные данные	11
2.1.3 Месячные данные	12
2.2 Выбор моделей прогнозирования рядов нижнего уровня	12
2.3 Группировка рядов третьего уровня по регионам, по типу и по метрике расстояния	14
2.4 Сравнение иерархических моделей	15
Заключение	19
Список литературы	21
Приложение А Визуализация временных рядов с трехуровневой структурой	22
Приложение Б Сравнение иерархических моделей	25

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

ний 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

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 агрегированных 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 оценкой 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 агрегированных рядов, существуют более сложные методы получения оптимальных комбинаций прогнозов, 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.

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.

Forward modelling: with noise properties we can predict the Sampling Distribution (the probability for a general set of data)

In easy cases, the effect of the prior is simple. As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes.

Rule of thumb: if changing your prior to another reasonable one changes the answers significantly, you need more data

Bayesians claim that the parameters are random so that their credible interval is a valid probability argument, though it also depends on the prior, which is usually hard to obtain.

When the likelihood and prior is complicated, the inference has to rely on the MCMC sampling, which can be really slow in most of the real-world cases.

The biggest controversy about Bayesian inference is that you must quantify your prior knowledge about the question at hand. This makes it possible to actually influence your results, either accidentally or on purpose.

Shrinkage is implicit in Bayesian inference and penalized likelihood inference, and explicit in James–Stein-type inference. In contrast, simple types of maximum-likelihood and least-squares estimation procedures do not include shrinkage effects, although they can be used within shrinkage estimation schemes.

Stein's paradox, in decision theory and estimation theory, is the phenomenon that when three or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (that is, having lower expected mean squared error) than any method that handles the parameters separately.

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

Stein's paradox **modern generalization, the Bayesian hierarchical model**.

Bayesian hierarchical model can improve overall estimation accuracy, thereby improving our confidence in the assessment results, especially for standard compliance assessment of waters with **small sample sizes.**

1.2 Взвешенные прогнозы

The assumption upon which many of these models are built on, is that by grouping series that behave in a similar way, the idiosyncratic errors within groups will tend to offset each other while the more relevant individual dynamics will be retained to be modelled.

Key idea: forecast reconciliation

- Ignore structural constraints and forecast every series of interest independently.
- Adjust forecasts to impose constraints.

Existing methods:

- Bottom-up
- Top-down
- Middle-out

An “optimal combination” approach can be advanced by proposing two new estimators based on WLS.

Both now implemented in the hts package

Y_t : observed aggregate of all series at time t .

$Y_{X,t}$: observation on series X at time t.

For the hierarchical structure we can write:

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

or in more compact notation, where y_t is an n -dimensional vector of all the observations in the hierarchy at time t , S is the summing matrix.

$$y_t = Sb_t$$

b_t : m -dimensional vector of all series at bottom level in time t.

$$\hat{Y}_n(h) = S\beta_n(h) + e_h$$

$\hat{Y}_n(h)$ - a vector of initial h -step forecasts, made at time n , stacked in same order as Y_t .

$$\beta_n(h) = E[B_{n+h}|Y_1, \dots, Y_n]$$

e_h has zero mean and covariance Σ_h .

$$\tilde{Y}_n(h) = S\hat{\beta}_n(h) = SP\hat{Y}_n(h) = S(S'\Sigma_h^+S)^{-1}S'\Sigma_h^+\hat{Y}_n(h)$$

$\tilde{Y}_n(h)$ - revised forecasts, $\hat{Y}_n(h)$ - initial forecasts

Σ_h^+ - generalized inverse of Σ_h .

Optimal $P = (S'\Sigma_h^+ S)^{-1} S'\Sigma_h^+$

Problem: Σ_h hard to estimate.

$$\tilde{Y}_n(h) = S(S'\Sigma_h^+ S)^{-1} S'\Sigma_h^+ \hat{Y}_n(h)$$

– OLS

$e_{B,h}$ is the forecast error at bottom level.

Assume $e_h \approx Se_{B,h}$ then $(S'\Sigma^+ S)^{-1} S'\Sigma^+ = (S'S)^{-1} S'$

$$\tilde{Y}_n(h) = S(S'S)^{-1} S' \hat{Y}_n(h)$$

– Rescaling

Suppose we approximate Σ_h by its diagonal: $\Lambda = \text{diag}(\hat{\Sigma}_1)^{-1}$, which contain inverse one-step ahead in-sample forecast error variances.

$$\tilde{Y}_n(h) = S(S'\Lambda S)^{-1} S' \Lambda \hat{Y}_n(h)$$

– Averaging

If the bottom level error series are approximately uncorrelated and have similar variances, then Λ is inversely proportional to the number of series contributing to each node: $\Lambda = \text{diag}(S \times 1)^{-1}$ is the inverse row sums of S

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 2.1

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, являющимися 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 аддитивности²:

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.

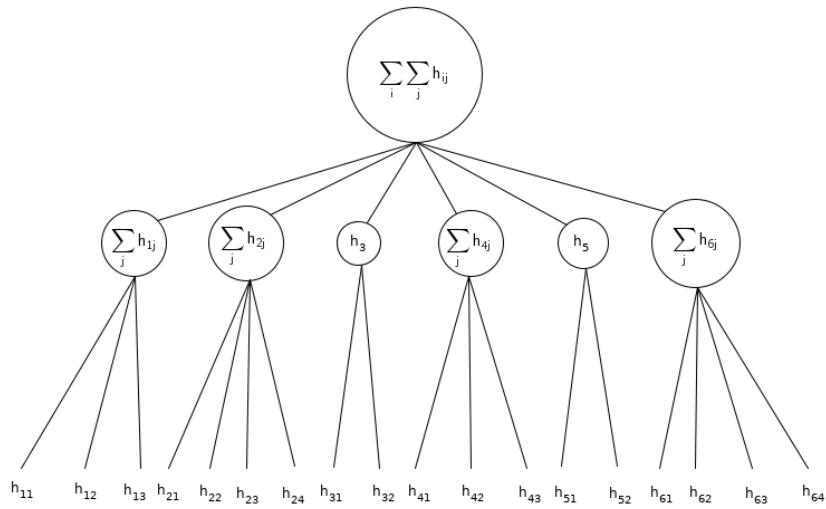


Рисунок 2.1 — Структура временных рядов, необходимых для исследования

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

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 А.

2.1.1 Квартальные данные

Квартальные данные³ - ряды ВДС по 28 странам Европейского союза (включая Великобританию) в разбивке по основным отраслям⁴:

- 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;

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

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

- 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.

2.1.2 Квартальные сезонно сглаженные данные

Квартальные сезонно сглаженные данные⁵ - 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 4⁶.

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%.

⁵FRED: Economic Data. URL: <https://fred.stlouisfed.org/>

⁶Алгоритм, 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.

2.1.3 Месячные данные

Месячные данные⁷ - 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 первого набора разрабатываются ряды только по основным классам и отдельным причинам смерти, имеющим наибольший вес. Также в 2011 году методика разработки показателя была пересмотрена, чтобы соответствовать Международной статистической классификации⁸.

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:

- 1) 'РО' - число рожденных;
- 2) 'УБ' - число умерших из-за болезней (болезней органов дыхания, органов пищеварения, системы кровообращения, инфекционных и паразитарных болезней, новообразований);
- 3) 'УУ' - число умерших по причине убийства и самоубийства;
- 4) '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

⁷ЕМИСС: государственная статистика: Официальные статистические показатели. URL: <https://www.fedstat.ru/>

⁸Демографический ежегодник России: методические пояснения. URL: http://www.gks.ru/bgd/reg1/B17_16/Main.htm

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.⁹

Для выбора параметров в моделях применяется кросс-валидация со скользящим окном 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.

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	$\lambda = 1, T = A$	$\lambda = 1, T = A$	$\lambda = 1, T = Ad$

Ширина окна для каждого из набора данных подбиралась в соответствии длинной ряда и горизонтом прогнозирования в два года таким образом, чтобы при проведении перекрестной проверки с шагом в один год получалось не менее пяти подвыборок. Соответственно для квартальных, сезонно сглаженных и месячных данных имеем:

⁹Автоматический перебор параметров модели осуществляется с помощью функций R 'auto.arima', 'ets' и 'tbats' соответственно

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

Различия в ширине окна позволяют сравнить 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.

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 (МЕ) 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 позволит 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 настолько большое число компонент излишне, поскольку приводит к тому, что ошибки прогнозов каждого ряда накапливаются, 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 двуми этими эффектами.

Очевидно, что можно сгруппировать ряды по территориальному признаку (по странам, штатам или регионам) или по типам (для ВВП по отраслям и для естественного прироста отдельно ряды по рождаемости и причинам смерти). Аддитивность позволяет получить ряды второго уровня просто просуммировав ряды, входящие в одну группу.

Альтернативным способом группировки будет кластеризация нормированных рядов по метрике евклидова расстояния с помощью алгоритма иерархической кластеризации, реализованного 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 А.3.

Очевидно, 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:

- для выполнения перекрестной проверки со скользящим окном с шагом в один год каждый из наборов данных делится на подвыборки: 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 для всех трех наборов данных (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.

В результате описанной процедуры получаются таблички, с помощью которых можно

наглядно увидеть, как изменился прогноз агрегированного ряда при использовании моделей, учитывающих многоуровневую структуру данных (Приложение Б).

Анализируя полученные показатели и визуальное представление анализируемых наборов данных (Приложение А) можно сделать следующие выводы:

Вывод 1:

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 А.2 где ряды группируются по типу). Соответственно для некоторых моделей хорошие прогнозы имела большая по размеру группа, а для некоторых меньшая.

Вывод 2:

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 отличающееся 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 ис-

пользовать 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 третья часть рядов прогнозируется по подходящей модели, а четверть по распределению совпадает с некачественным агрегированным прогнозом.

Вывод 3:

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

При анализе рисунка А.2 группировка по типу приводит к тому, что в большинстве слу-

чаев 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. — Т. 53. — С. 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. — Т. 25, № 1. — С. 146—166.
4. Clark J. S. Uncertainty and variability in demography and population growth: a hierarchical approach // *Ecology*. — 2003. — Т. 84, № 6. — С. 1370—1381.
5. Cobb M. Forecasting Economic Aggregates Using Dynamic Component Grouping. — 2017.
6. Cobb M. Joint forecast combination of macroeconomic aggregates and their components. — 2017.
7. Diebold F. X., Pauly P. The use of prior information in forecast combination // *International Journal of Forecasting*. — 1990. — Т. 6, № 4. — С. 503—508.
8. Duncan G. T., Gorr W. L., Szczypula J. Forecasting analogous time series. — 2001.
9. Fox D. R. Concepts and Methods of the U.S. National Income and Product Accounts. — Bureau of Economic Analysis (BEA), 2017.
10. Gelman A. Multilevel (hierarchical) modeling: what it can and cannot do // *Technometrics*. — 2006. — Т. 48, № 3. — С. 432—435.
11. Green K. C., Armstrong J. S. Structured analogies for forecasting // *International Journal of Forecasting*. — 2007. — Т. 23, № 3. — С. 365—376.
12. Hyndman R. J., Lee A. J., Wang E. Fast computation of reconciled forecasts for hierarchical and grouped time series // *Computational Statistics & Data Analysis*. — 2016. — Т. 97. — С. 16—32.
13. Katz A. J. An Overview of BEA's Source Data and Estimating Methods for Quarterly GDP. — 10th OECD-NBS Workshop on National Accounts, 2006.
14. Makridakis S., Hibon M. The M3-Competition: results, conclusions and implications // *International journal of forecasting*. — 2000. — Т. 16, № 4. — С. 451—476.
15. McNeish D., Wentzel K. R. Accommodating small sample sizes in three-level models when the third level is incidental // *Multivariate behavioral research*. — 2017. — Т. 52, № 2. — С. 200—215.
16. Moyer B. C., Thompson S. Gross Domestic Product by State Estimation Methodology. — 2017.
17. Optimal combination forecasts for hierarchical time series / R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, H. L. Shang // *Computational Statistics & Data Analysis*. — 2011. — Т. 55, № 9. — С. 2579—2589.
18. 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. — Т. 26, № 2. — С. 330—343.
19. Shang H. L., Smith P. W. Grouped time-series forecasting with an application to regional infant mortality counts. — 2013.

20. Stegmüller D. How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches // American Journal of Political Science. — 2013. — T. 57, № 3. — C. 748—761.
21. Tobias J. L. Forecasting output growth rates and median output growth rates: A hierarchical Bayesian approach // Journal of Forecasting. — 2001. — T. 20, № 5. — C. 297—314.
22. Weiss C. Essays in Hierarchical Time Series Forecasting and Forecast Combination. — 2018.
23. Zellner A., Hong C. Forecasting international growth rates using Bayesian shrinkage and other procedures // Journal of Econometrics. — 1989. — T. 40, № 1. — C. 183—202.

Приложение А Визуализация временных рядов с трехуровневой структурой

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

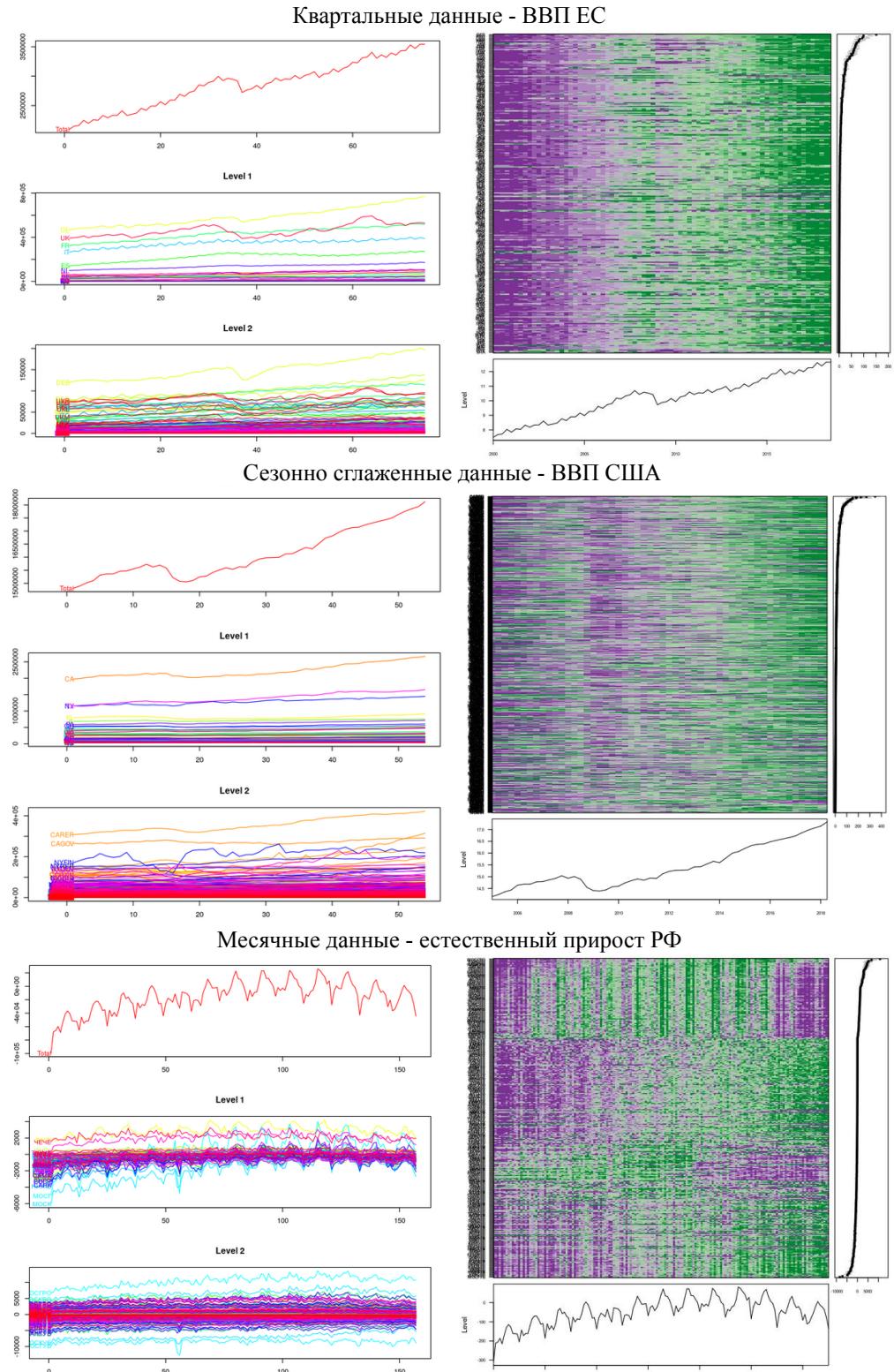


Рисунок А.2 — Временные ряды сгруппированные по территориальному признаку и по типу

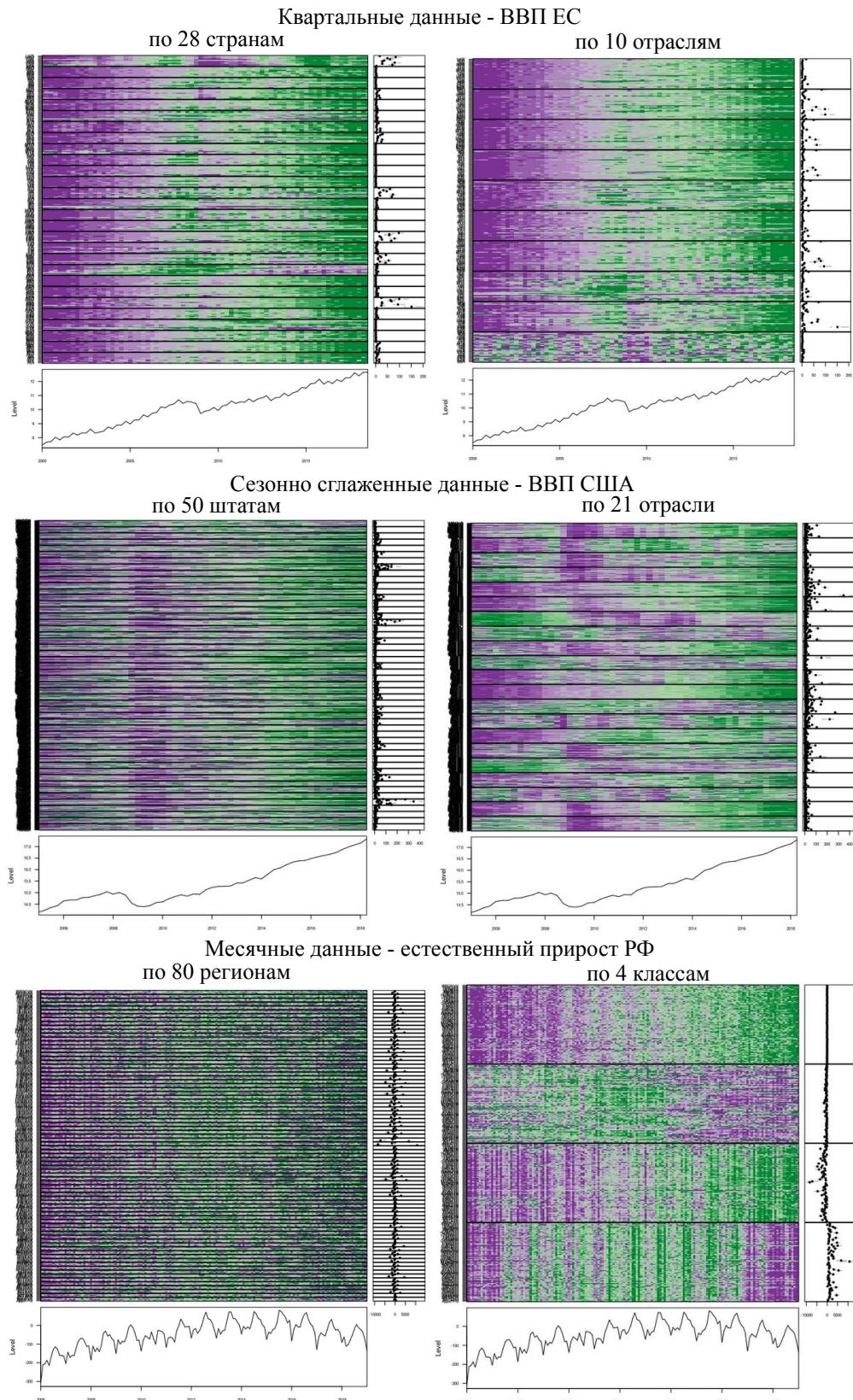
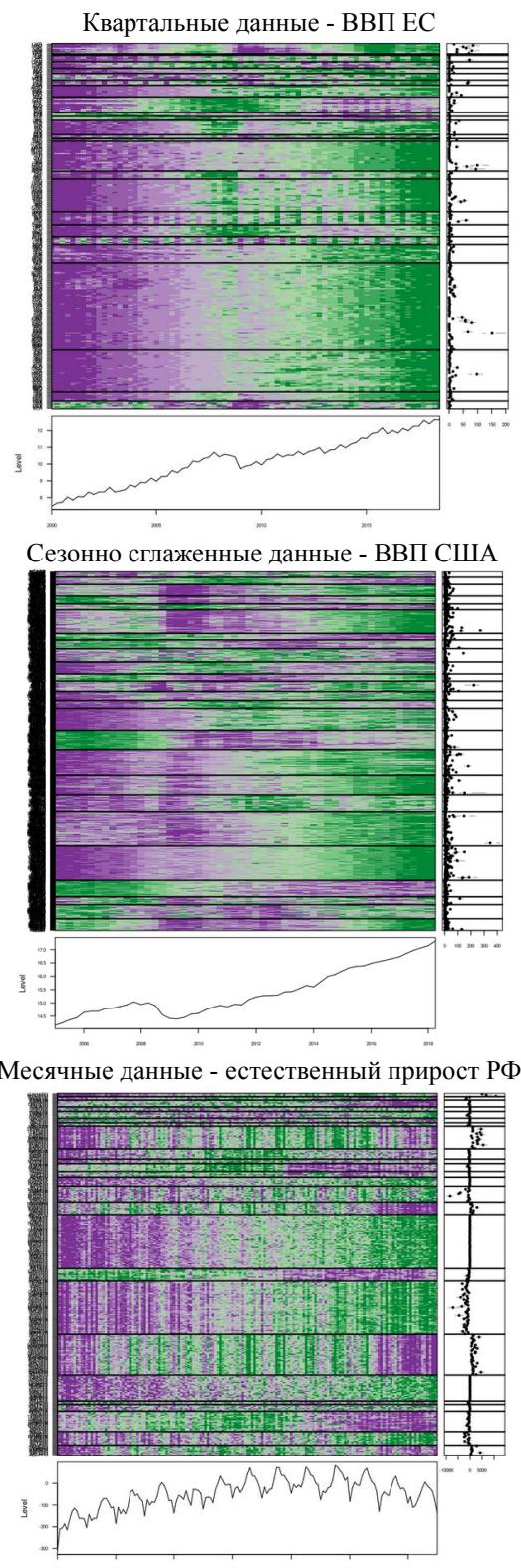


Рисунок А.3 — Временные ряды сгруппированные по метрике расстояния



Приложение Б Сравнение иерархических моделей



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 /