

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ



Душан Пантелић

АНАЛИЗА УТИЦАЈА ХИЈЕРАРХИЈСКОГ
УСКЛАЂИВАЊА НА ТАЧНОСТ
ПРЕДВИЂАЊА ВРЕМЕНСКИХ СЕРИЈА
МОДЕЛИМА МАШИНСКОГ УЧЕЊА

мастер рад

Београд, 2026.

Ментор:

др Младен НИКОЛИЋ, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Бојана МИЛОШЕВИЋ, ванредни професор
Универзитет у Београду, Математички факултет

др Мирјана МАЉКОВИЋ РУЖИЧИЋ, доцент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Наслов мастер рада: Анализа утицаја хијерархијског усклађивања на тачност предвиђања временских серија моделима машинског учења

Резиме: Управљање ланцима снабдевања захтева прецизно предвиђање потражње на више нивоа агрегације, што је значајно отежано изразитом испрекиданошћу временских серија. Циљ овог рада био је да се испита утицај различитих метода хијерархијског усклађивања на прецизност базних прогноза и квантификује губитак математичке кохерентности приликом хеуристичког третмана негативних вредности. Истраживање је спроведено на скупу од 20.000 временских серија, уз евалуацију једанаест алгоритама (од статистичких модела до савремених вишенаменских претренираних мрежа). Базне прогнозе су подвргнуте структурном и темпоралном усклађивању (*OLS*, *WLS_Struct*, *WLS_Var*, *MinT_Shrink*), уз упоредну анализу строге оптимизације (QP) и хеуристичког одсецања негативних вредности на нулу. Резултати показују да усклађивање не побољшава универзално тачност система. Модели дубоког учења (*DeepAR*, *N-HITS*) и вишенаменски претренирани модели (*Chronos*, *TimesFM*) остварују најбоље перформансе, док детерминистичко усклађивање (*OLS*) драстично деградира њихов прецизан локални сигнал. Показано је да темпорално усклађивање функционише примарно као пост-хок регуларизатор који значајно ублажава грешке изразито лоших модела (пад грешке до 60% код *LightGBM*-а), али истовремено квари сигнале прецизних алгоритама. Такође, доказано је да хеуристичко одсецање негативних прогноза доводи до просечног нарушавања кохерентности од 6.00%, што се применом методе *WLS_Var* своди на занемарљивих 0.16% у просеку, са максималних 0.5% појединачно по моделима. Закључак указује да хеуристичко одсецање негативних вредности, упарено са методама које поштују историјску дисперзију, представља бољу праксу од нестабилних QP решавача, пружајући директне смернице за пројектовање робусних предиктивних система у разматраном домену.

Кључне речи: хијерархијске временске серије, хијерархијско усклађивање, предвиђање временских серија, испрекидана потражња, машинско учење, дубоко учење, вишенаменски претренирани модели, статистички модели

Садржај

1	Увод	1
1.1	Контекст и значај проблема и области	1
1.2	Дефиниција истраживачког проблема	3
1.3	Примарни циљеви рада	4
1.4	Кључни резултати	5
2	Теоријске основе	7
2.1	Основни концепти предвиђања временских серија	7
2.2	Хијерархијске и груписане временске серије	8
2.3	Методе за усклађивање предвиђања	11
3	Дизајн истраживања и методологија	22
3.1	Карактеристике и припрема скупа података	22
3.2	Поставка и процес спровођења експеримента	28
4	Резултати и дискусија	37
4.1	Анализа перформанси неусклађених предвиђања	37
4.2	Упоредна анализа метода за усклађивање	39
4.3	Валидација и нарушавање хијерархијске кохерентности	45
4.4	Дискусија добијених резултата	47
4.5	Идентификација најбољих пракси за дати проблем	49
5	Закључак	51
5.1	Синтеза кључних налаза истраживања	51
5.2	Ограничења спроведене анализе	52
5.3	Препоруке за будућа истраживања и практичну примену	53
	Библиографија	54

Глава 1

Увод

Предвиђање (прогнозирање) временских серија (енг. time series forecasting) представља фундаменталну област машинског учења и примењене статистике, која за циљ има екстракцију образаца из секвенцијалних података. Иако ови концепти налазе примену у разним доменима, од обраде сигнала до метеорологије, њихов највећи практични утицај је у домену пословних система и ланаца снабдевања [14]. Данашње свакодневно пословно окружење карактерише изузетна динамичност, која је резултат непредвидивости и све веће сложености тржишта. Способност организација да доносе правовремене одлуке на основу предвиђања будућих догађаја директно утиче на њихову стабилност и конкурентност. Међутим, реални подаци који описују пословне системе ретко су изоловане временске серије. Наиме, подаци унутар једне организације готово увек поседују инхерентну хијерархијску или групну структуру. Усклађивање ових процеса кроз све нивое одлучивања представља један од највећих изазова [10].

1.1 Контекст и значај проблема и области

Предвиђање у системима са хијерархијама захтева предвиђање догађаја на различитим нивоима агрегације. Хијерархије се могу поделити на две основне димензије:

1. **Структурне:** Представљају просторну или логичку агрегацију, односно декомпозицију података у фиксном временском интервалу. На пример, у глобалном ланцу снабдевања, укупна светска тражња се деком-

понује на регионе, државе, дистрибутивне центре, и на крају на појединачне артикле (енг. stock keeping unit – SKU).

2. **Временске:** Подразумевају агрегацију података високе учесталости (нпр. дневне трансакције) у ниже фреквенције (недељне, месечне, годишње серије), што омогућава моделовање краткорочних оперативних осцилација и дугорочних стратешких трендова [3].

Традиционално се овом проблему приступа применом хеуристичких метода. Приступ „од дна ка врху” (енг. bottom-up) предвиђа само најниже нивое који се потом сабирају, али је изузетно осетљив на висок ниво шума карактеристичан за податке на ниском нивоу [6]. Са друге стране, приступ „од врха ка дну” (енг. top-down) користи стабилност агрегираних података на вишим нивоима [8], али често не успева да прецизно декомпонује информације, губећи битне локалне карактеристике података [2]. Независно генерисане прогнозе на различитим нивоима су по правилу математички некохерентне (збир нижих нивоа не одговара прогнози вишег нивоа), што може довести до контрадикторних пословних одлука.

Као примарно решење овог проблема развијена је парадигма оптималног усклађивања (енг. optimal reconciliation) [9], са методом минималног трага (енг. minimum trace – MinT) као тренутним стандардом [22]. Иако је иницијални мотив за развој ових метода била административна кохерентност, новија сазнања указују на то да усклађивање заправо функционише као механизам регуларизације [12]. Приморавањем модела са различитих нивоа да „комуницирају”, усклађивање омогућава пренос информација, стабилност и ниска дисперзија са виших нивоа хијерархије спуштају се на нижи ниво, док се карактеристични сигнали са дна пењу ка врху. Управо ова информациона размена, а не пука математичка кохерентност, представља потенцијал за унапређење предиктивних система.

Досадашња литература је овај домен претежно анализирала кроз употребу класичних модела машинског учења и статистичких модела. Скорашњи напредак архитектура дубоког учења [13, 18], као и појава тзв. вишенамених претренираних (енг. foundation) модела обучених на великим скуповима података [1, 7], захтевају потпуно нову евалуацију у контексту хијерархијског усклађивања.

1.2 Дефиниција истраживачког проблема

Суштина истраживачког проблема лежи у промени парадигме: усклађивање независних прогноза се у овом раду не посматра искључиво као захтев за изједначавање збирова, већ као механизам за побољшање предвиђања кроз размену информација између нивоа хијерархије. Проблем се огледа у проналажењу оптималног математичког модела који ће ово омогућити. Метод минималног трага постиже размену информација ослањајући се на апроксимацију коваријационе матрице грешака ван узорка, помоћу података из узорка (резидуала или историјских вредности). Међутим, примена ове методологије на реалне податке открива неколико изазова којима се овај рад првенствено бави:

- 1. Проблем апроксимације структуре грешака:** Иако увођење пуне узорачке коваријационе матрице (основна варијанта методе минималног трага) теоријски омогућава најпотпунији пренос информација, њена примена на хијерархије са великим бројем временских серија је статистички и нумерички нестабилна. Када број временских серија (20.000) драстично превазилази број расположивих историјских опажања (~ 60), пуна узорачка коваријациона матрица постаје непотпуног ранга и сингуларна, што онемогућава њену инверзију [20]. Услед ових ограничења, варијанте са пуном узорачком коваријансом су искључене из анализе. Истраживање се фокусира на евалуацију преосталих стабилних алтернатива: OLS (метод обичних најмањих квадрата) и WLS_Struct (метод пондерисаних најмањих квадрата заснован на структури хијерархије) које нестабилност заобилазе ослањајући се на геометрију и топологију хијерархије без анализе грешака, WLS_Var (метод пондерисаних најмањих квадрата заснован на дисперзији) која уводи само дијагоналне дисперзије резидуала, као и напредне регуларизационе технике MinT_Shrink, која поступком математичког „скупљања” (енг. shrinkage) стабилизује пуну коваријациону матрицу.
- 2. Сукоб математичке кохерентности и физичких ограничења:** Усклађивање често резултује негативним прогнозама, што је у контексту прогнозирања потражње физички немогуће и доводи до деградације евалуационих метрика. Формална ограничења ненегативности се

успешно имплементирају кроз квадратно програмирање, односно оптимизацију, што за последицу има прелазак са аналитичког решења на итеративно тражење оптимума. Овај прелазак уводи озбиљне нумеричке изазове и рачунарску комплексност у условима високе димензионалности и ограниченог историјског узорка. Иако ова појава зависи од специфичне топологије података и резидуала, она у оквиру нашег експеримента узрокује да итеративни поступци за емпиријске методе (*WLS_Var* и *MinT_Shrink*) доживљавају нумеричку нестабилност услед лоше условљености матрица израчунатих из резидуала. Насупрот њима, детерминистичке методе (*OLS* и *WLS_Struct*) успешно конвергирају на нашим подацима. С обзиром на то да овај рад евалуира перформансе модела и са формалним ограничењима и кроз стратегију хеуристичког одсецања негативних вредности на нулу (енг. *clipping to zero*), отвара се истраживачко питање: Како избор између строге математичке кохерентности и накнадних интервенција утиче на прецизност прогноза на циљаном (базном) нивоу и да ли се задржавају бенефити преноса информација?

3. **Хетерогеност структура грешака:** Различите класе модела генеришу суштински различите обрасце резидуала. Истраживачки изазов је разумевање интеракције ове хетерогености унутар метода за усклађивање и како утичу на коначан резултат.

1.3 Примарни циљеви рада

Да би се успешно одговорило на дефинисани истраживачки проблем, постављени су следећи примарни циљеви:

1. **Системска евалуација примењивих варијанти MinT:** Спровести свеобухватну компаративну анализу четири методе усклађивања (*OLS*, *WLS_Struct*, *WLS_Var* и *MinT_Shrink*) на структурним и временским хијерархијама.
2. **Анализа утицаја ограничења ненегативности:** Испитати како увођење формалних ограничења ненегативности (за *OLS* и *WLS_Struct*), у поређењу са стратегијом накнадног хеуристичког одсецања негативних вредности на нулу утиче на метрике. Циљ је демонстрирати да се

предности размене информација кроз нивое могу задржати чак и када је строга математичка кохерентност делимично нарушена ради очувања пословне логике и рачунарских перформанси.

3. Компаративна евалуација четири фамилије предиктивних модела:

- Статистички модели (*ARIMA*, *Holt-Winters*, *Prophet*).
- Класични модели машинског учења (*LightGBM*, *RandomForest*).
- Модели дубоког учења (*DeepAR*, *N-HITS*).
- Вишенаменски претренирани модели за временске серије (*Chronos*, *TimesFM*).

4. Идентификација најбољих пракси: Синтетизовати резултате у јасне смернице које ће допринети лакшем избору оптималне комбинације модела, методе усклађивања и стратегије третмана негативних прогноза у реалним пословним системима.

1.4 Кључни резултати

Темељна експериментална евалуација спроведена у овом раду довела је до неколико увида:

- **Супериорност савремених архитектура:** На посматраном скупу података, вишенаменски претренирани модели и вероватносне неуронске мреже вишеструко надмашују класичне алгоритме машинског учења. Стабла одлучивања са стандардном функцијом циља показују велику пристрасност и нису адекватна за серије са великим бројем нула.
- **Ограничења хијерархијског усклађивања:** Показано је да усклађивање не побољшава аутоматски прецизност базних прогноза. Детерминистички приступи најчешће деградирају прецизне локалне сигнале, док је квалитет добрих модела могуће сачувати искључиво применом метода које интелигентно скалирају корекције.
- **Темпорално усклађивање као регуларизатор:** Усклађивање кроз време не доноси стварну предиктивну вредност већ оптималним алго-

ритмима. Оно делује примарно као механизам пост-хок регуларизације који ублажава екстремне грешке модела лошијих перформанси.

- **Оправданост инжењерског компромиса:** Показано је да строга математичка оптимизација за спречавање негативних прогноза није рачунарски одржива у високодимензионалним корпоративним системима. Једноставно хеуристичко одсецање негативних вредности представља бољи правац, јер обезбеђује физички смислене прогнозе уз занемарљив губитак агрегационе кохерентности.
- **Утицај дизајна хијерархије:** Комплекснија структура хијерархије уноси додатни шум у систем. Редуковање броја агрегационих нивоа значајно ублажава деградацију базних прогноза услед смањеног броја алгебарских ограничења.

Остатак рада је структуриран у четири поглавља. У поглављу 2 изложене су теоријске основе хијерархијских временских серија и методе за њихово усклађивање. Поглавље 3 описује експерименталну поставку, коришћени скуп података, архитектуре модела и метрике евалуације. У поглављу 4 приказани су резултати експеримената уз пратећу дискусију, док су у поглављу 5 изведени главни закључци и дате смернице за даља истраживања.

Глава 2

Теоријске основе

Ово поглавље поставља формални математички и статистички оквир неопходан за разумевање проблема предвиђања у сложеним системима. Први део поглавља уводи основне концепте временских серија. Затим се дефинишу алгебарске структуре хијерархијских и груписаних временских серија. Коначно, трећи део поглавља доноси детаљан теоријски преглед метода за усклађивање предвиђања.

2.1 Основни концепти предвиђања временских серија

Временска серија се математички дефинише као низ опажања посматраних у узастопним, најчешће једнако удаљеним временским интервалима. Нека је y_t вредност посматране променљиве у тренутку t . Временска серија дужине T може се представити као секвенца $\mathbf{y} = (y_1, y_2, \dots, y_T)$. Циљ предвиђања (прогнозирања) јесте процена будућих вредности, које означавамо са $\hat{y}_{T+h}$ за $h = 1, 2, \dots, H$, где h представља хоризонт предвиђања, на основу историјских података и потенцијалних екстерних фактора [10].

Иако класична теорија временских серија често посматра податке кроз експлицитне декомпозиције на тренд, сезоналност и шум, у савременим предиктивним системима, моделима се најчешће прослеђује сирова, недекомпована серија. Без обзира на то да ли припадају фамилији статистичких метода, алгоритмима машинског учења, или архитектурама дубоког учења, од модела се очекује да самостално екстрахују релевантне темпоралне инфор-

мације директно из оригиналног сигнала.

У контексту ланаца снабдевања, посебан изазов представљају подаци на најнижим нивоима агрегације који некада попримају облик **испрекиданих временских серија** (енг. *intermittent time series*). Испрекидане серије карактерише велики број нултих вредности (периоди без икакве потражње) испрекидан повременим, често варијабилним позитивним вредностима када се потражња заправо деси.

Управо ово својство чини предвиђање серија изузетно тешким и подложним високом нивоу шума, што потенцијално можемо решити позајмљивањем стабилнијих информација са виших нивоа агрегације кроз процес хијерархијског усклађивања (где се скуп индивидуалних серија y_t третира као јединствен систем више променљивих y_t), како би се побољшале предикције на целом систему.

Да бисмо ово приказали, посматрајмо једноставан пример приказан у табели 2.1. Потражња за специфичним делом у три различита града (базне серије) је изразито испрекидана. Међутим, њихов збир на регионалном нивоу (агрегирана серија) формира стабилан сигнал без нултих вредности.

Табела 2.1: Пример агрегације испрекиданих серија у непрекидну серију.

Серија	t_1	t_2	t_3	t_4	t_5	t_6
град А (базни ниво)	0	2	0	0	0	1
град Б (базни ниво)	0	0	1	0	2	0
град В (базни ниво)	1	0	0	3	0	0
регион (збир)	1	2	1	3	2	1

2.2 Хијерархијске и груписане временске серије

У реалним пословним системима, временске серије нису изоловани ентитети, већ су међусобно повезане строгим правилима агрегације. Ове структуре се генерално деле на хијерархијске (где постоји строги редослед рашчлањивања, нпр. држава $\rightarrow$ регион $\rightarrow$ град) и груписане (где се подаци могу агрегирати по више независних атрибута, нпр. агрегација по географској локацији или алтернативно по типу производа).

Математичка формулација структурних хијерархија

Нека систем садржи укупно n временских серија, од којих се m серија налази на најнижем (базном) нивоу, док преосталих $k = n - m$ серија представљају агрегиране вредности на вишим нивоима.

Вектор свих опажања у тренутку t , означен са $\mathbf{y}_t$ (димензија $n \times 1$), може се партиционисати на вектор агрегираних серија $\mathbf{a}_t$ (димензија $k \times 1$) и вектор базних серија $\mathbf{b}_t$ (димензија $m \times 1$):

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{a}_t \\ \mathbf{b}_t \end{bmatrix} \quad (2.1)$$

Кључни алгебарски елемент који дефинише структуру овог система јесте **матрица сабирања** $\mathbf{S}$ (енг. *summing matrix*) димензија $n \times m$. Ова матрица пресликава базне серије у све остале серије у систему. Коришћењем матрице $\mathbf{S}$, цео систем се може изразити искључиво преко базних серија:

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t \quad (2.2)$$

Матрица $\mathbf{S}$ се може партиционисати као:

$$\mathbf{S} = \begin{bmatrix} \mathbf{A} \\ \mathbf{I}_m \end{bmatrix} \quad (2.3)$$

где је $\mathbf{A}$ агрегациона матрица димензија $k \times m$ која дефинише како се базне серије сабирају да би се добили виши нивои, док је $\mathbf{I}_m$ јединична матрица димензија $m \times m$ која пресликава базне серије саме на себе. Елементи матрице $\mathbf{A}$ су типично бинарни (0 или 1), где 1 означава да базна серија припада одговарајућем сумирању.

Да би се илустровала моћ и флексибилност ове формулације, у наставку су приказана два карактеристична примера: један за стриктну хијерархију и један за груписане серије.

Пример 1: Стриктна хијерархија (стабло)

Посматрајмо систем са три нивоа, где је укупна продаја државе (највиши ниво, y_T) декомпонована на два региона: север (y_S) и југ (y_J). Даље, регион север чине два града (y_{S1}, y_{S2}), а регион југ такође два града (y_{J1}, y_{J2}). Ово је класична структура стабла где су нижи нивои стриктни подскупови виших.

У овом систему имамо $m = 4$ базне серије (градови) и укупно $n = 7$ серија. Матрица сабирања $\mathbf{S}$ димензија 7×4 формира се на следећи начин:

$$\begin{bmatrix} y_T \\ y_S \\ y_J \\ y_{S1} \\ y_{S2} \\ y_{J1} \\ y_{J2} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{S}} \begin{bmatrix} y_{S1} \\ y_{S2} \\ y_{J1} \\ y_{J2} \end{bmatrix} \quad (2.4)$$

Горња три реда представљају агрегациону матрицу $\mathbf{A}$, док су доња четири реда јединична матрица $\mathbf{I}_4$.

Пример 2: Груписане временске серије

За разлику од стриктне хијерархије, подаци се често могу груписати по више независних атрибута. Претпоставимо да компанија прати продају по географском региону (регион А и регион Б) и алтернативно по категорији производа (производ Х и производ Y).

Најнижи (базни) ниво чине све комбинације ових атрибута $(y_{AX}, y_{AY}, y_{BX}, y_{BY})$, што даје $m = 4$. Виши нивои чине: укупна продаја (y_T) , укупна продаја по регионима (y_A, y_B) и укупна продаја по производима (y_X, y_Y) . Укупно имамо $n = 9$ серија. Матрица сабирања $\mathbf{S}$ сада изгледа овако:

$$\begin{bmatrix} y_T \\ y_A \\ y_B \\ y_X \\ y_Y \\ y_{AX} \\ y_{AY} \\ y_{BX} \\ y_{BY} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{S}} \begin{bmatrix} y_{AX} \\ y_{AY} \\ y_{BX} \\ y_{BY} \end{bmatrix} \quad (2.5)$$

Док редови за регионе врше класично сабирање суседних елемената (као у примеру 1), редови за производе $(y_X$ и $y_Y)$ врше унакрсно сабирање (нпр.

$y_X = y_{AX} + y_{BX}$). Целокупна комплексност било стриктне хијерархије, било унакрсног груписања, садржана је унутар структуре матрице $\mathbf{S}$.

Временске хијерархије

Поред структурне агрегације, временске серије се могу посматрати кроз призму временских хијерархија [3]. Ако претпоставимо да располажемо подацима високе фреквенције, ова опажања се могу агрегирати у ниже фреквенције. Временска хијерархија омогућава да се шум у високофреквентним подацима ублажи кроз агрегацију, док нискофреквентни подаци олакшавају моделирање дугорочних трендова.

Концепт изоморфизма између структурних и временских хијерархија може се приказати на примеру агрегације високофреквентних **месечних података у нискофреквентну кварталну серију** (што уједно одговара и временској агрегацији примењеној у овом истраживању).

Нека су y_{M1}, y_{M2}, y_{M3} опажања за три узастопна месеца (базне серије у временском домену), а y_{Q1} њихова укупна квартална вредност (агрегирана серија). Временски вектор система за овај квартал је $\mathbf{y}_{temp} = \begin{bmatrix} y_{Q1} & y_{M1} & y_{M2} & y_{M3} \end{bmatrix}^T$.

Припадајућа матрица сабирања $\mathbf{S}_{temp}$ димензија 4×3 формира се по идентичном принципу:

$$\begin{bmatrix} y_{Q1} \\ y_{M1} \\ y_{M2} \\ y_{M3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{M1} \\ y_{M2} \\ y_{M3} \end{bmatrix} \quad (2.6)$$

Кроз ове примере јасно се уочава да без обзира на то да ли се подаци агрегирају просторно или темпорално, алгебарска структура проблема (релација $\mathbf{y} = \mathbf{Sb}$) остаје потпуно иста. Ово својство омогућава примену идентичних алгоритама за оптимално усклађивање како на просторној, тако и на временској димензији.

2.3 Методе за усклађивање предвиђања

Када се за сваку серију у хијерархији независно генерише предвиђање, добијамо вектор неусклађених прогноза $\hat{\mathbf{y}}_t$. Због независне природе предик-

тивних процеса, ове прогнозе готово никада не задовољавају хијерархијску структуру, односно:

$$\hat{\mathbf{y}}_t \neq \mathbf{S}\tilde{\mathbf{b}}_t \quad (2.7)$$

Овај недостатак кохерентности решава се приступом усклађивања.

Општи оквир усклађивања и традиционалне методе

Све линеарне методе усклађивања могу се представити као линеарне трансформације које пресликавају независне прогнозе у нове вредности које строго задовољавају правила хијерархијске агрегације. Комплетан вектор усклађених прогноза $\tilde{\mathbf{y}}_t$, који садржи ревидиране вредности за све серије на свим нивоима система, добија се применом следеће линеарне трансформације на вектор неусклађених прогноза $\hat{\mathbf{y}}_t$:

$$\tilde{\mathbf{y}}_t = \mathbf{S}\mathbf{P}\hat{\mathbf{y}}_t \quad (2.8)$$

Математички посматрано, ова трансформација се одвија у два логичка корака. Прво, матрица пресликавања $\mathbf{P}$ (димензија $m \times n$) узима информације из неусклађених прогноза са свих нивоа ($\hat{\mathbf{y}}_t$) и пресликава их у вектор кохерентних прогноза на основном (базном) нивоу. Затим, множењем са матрицом сабирања $\mathbf{S}$, те усклађене базне прогнозе се агрегирају. Коначни резултат ове операције ($\tilde{\mathbf{y}}_t$) јесте потпуно усклађен систем у коме су прогнозе на свим нивоима хијерархије математички конзистентне. Иако је у пракси базни ниво најчешће примарни циљ за евалуацију предиктивних метрика и оперативну реализацију, процес усклађивања унапређује и синхронизује цео систем.

Овај алгебарски оквир је универзалан [22]. Све традиционалне, хеуристичке методе усклађивања могу се представити специфичним избором елемената унутар матрице $\mathbf{P}$:

Приступ од дна ка врху:

Ова метода, која представља један од најстаријих приступа у хијерархијском прогнозирању, претпоставља да су прогнозе на најнижем нивоу једине релевантне, док се прогнозе генерисане на вишим нивоима потпуно игноришу [6]. Усклађене прогнозе виших нивоа добијају се искључиво сабирањем базних прогноза. Матрица $\mathbf{P}$ за ову методу дефинише се као:

$$\mathbf{P}_{BU} = \begin{bmatrix} \mathbf{0}_{m \times k} & \mathbf{I}_m \end{bmatrix} \quad (2.9)$$

где је $\mathbf{0}_{m \times k}$ нулти блок који множи и поништава све прогнозе са виших нивоа (укупно k серија), док јединична матрица $\mathbf{I}_m$ пропушта базне прогнозе непромењене.

Да бисмо ово приказали на примеру 1 из претходног поглавља (где имамо $n = 7$ серија и $m = 4$ града као базне серије), вектор свих неусклађених прогноза је $\hat{\mathbf{y}}_t = [\hat{y}_T, \hat{y}_S, \hat{y}_J, \hat{y}_{S1}, \hat{y}_{S2}, \hat{y}_{J1}, \hat{y}_{J2}]^T$. Матрица $\mathbf{P}_{BU}$ димензија 4×7 изгледа овако:

$$\mathbf{P}_{BU} = \left[\begin{array}{ccc|cccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \quad (2.10)$$

Множењем $\mathbf{P}_{BU}\hat{\mathbf{y}}_t$, прве три колоне (нулти блок) потпуно елиминишу прогнозе $\hat{y}_T, \hat{y}_S, \hat{y}_J$, остављајући само четири базне прогнозе које се затим преко матрице $\mathbf{S}$ сабирају навише. Главни недостатак овог приступа је очигледан: огромна количина информација са врха хијерархије се просто одбацује.

Приступ од врха ка дну:

Ова метода ради потпуно супротно: узима у обзир искључиво прогнозу са највишег (глобалног) нивоа, док игнорише све остале [8]. Глобална прогноза се затим дистрибуира на ниже нивое помоћу вектора историјских пропорција $\mathbf{p} = [p_1, p_2, \dots, p_m]^T$, где важи $\sum_{i=1}^m p_i = 1$. Матрица $\mathbf{P}$ за ову методу има облик:

$$\mathbf{P}_{TD} = \left[\begin{array}{c|c} \mathbf{p} & \mathbf{0}_{m \times (n-1)} \end{array} \right] \quad (2.11)$$

Ако ово применимо на исти пример 1 са 7 серија, и претпоставимо да смо израчунали вектор пропорција $\mathbf{p} = [p_{S1}, p_{S2}, p_{J1}, p_{J2}]^T$ (нпр. на основу просечне историјске продаје сваког града), матрица $\mathbf{P}_{TD}$ изгледа овако:

$$\mathbf{P}_{TD} = \left[\begin{array}{c|cccccc} p_{S1} & 0 & 0 & 0 & 0 & 0 & 0 \\ p_{S2} & 0 & 0 & 0 & 0 & 0 & 0 \\ p_{J1} & 0 & 0 & 0 & 0 & 0 & 0 \\ p_{J2} & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] \quad (2.12)$$

Када ову матрицу помножимо са $\hat{\mathbf{y}}_t$, она из целог система „извлачи” искључиво једну једину вредност, глобалну прогнозу целе државе $\hat{y}_T$ (преко прве колоне), множи је са пропорцијама, док нулти блок елиминише апсолутно све остале прогнозе (укључујући регионалне и градске). Недостатак овог при-

ступа је губитак специфичних локалних динамика, јер се све базне серије присиљавају да прате глобални образац.

Приступ из средине (енг. *middle-out*):

Ова метода представља хибрид претходне две [10]. Бира се један „средњи” ниво у хијерархији. Прогнозе са тог нивоа су једине које се задржавају; све прогнозе изнад тог нивоа се елиминишу и накнадно рачунају сабирањем (као код *bottom-up* методе), док се прогнозе испод тог нивоа елиминишу и рачунају дистрибуцијом преко пропорција (као код *top-down* методе).

Ако се вратимо на наш пример 1 ($n = 7$ серија: 1 држава, 2 региона, 4 града), претпоставимо да смо изабрали средњи ниво, **регионе** ($\hat{y}_S, \hat{y}_J$), као референтан. Прогнозе државе и свих 4 градова се одбацују. Потребан нам је вектор пропорција $\mathbf{p} = [p_{S1}, p_{S2}, p_{J1}, p_{J2}]^T$ који говори како се региони деле на градове. Важи правило да се градови на северу сабирају у један ($p_{S1} + p_{S2} = 1$), а градови на југу такође у један ($p_{J1} + p_{J2} = 1$).

Матрица $\mathbf{P}_{MO}$ за овај пример изгледа овако:

$$\mathbf{P}_{MO} = \left[\begin{array}{c|cc|cccc} 0 & p_{S1} & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{S2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{J1} & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{J2} & 0 & 0 & 0 & 0 \end{array} \right] \quad (2.13)$$

Када ову матрицу помножимо са $\hat{\mathbf{y}}_t$, она ради две ствари истовремено: нулти блок на почетку убија прогнозу државе ($\hat{y}_T$), блок на крају убија локалне прогнозе градова ($\hat{y}_{S1}, \dots$), док се из средњег дела узимају само регионалне прогнозе ($\hat{y}_S, \hat{y}_J$) и одмах множе пропорцијама p како би се спустиле на базни ниво градова.

Иако је флексибилнији, овај метод и даље пати од фундаменталног проблема: његова $\mathbf{P}$ матрица је препуна нула. Он одбацује велики део информација (игнорише прогнозе на свим нивоима осим на том једном изабраном) и што је још важније, тешко се може применити на груписане серије без строге хијерархије где не постоји јасан „средњи” ниво.

Кључни закључак који произилази из анализе ових матрица ($\mathbf{P}_{BU}, \mathbf{P}_{TD}, \mathbf{P}_{MO}$) јесте да традиционалне хеуристике драстично ограничавају проток информација јер намерно поништавају огроман део вектора $\hat{\mathbf{y}}_t$. То ствара математичку потребу за проналажењем **оптималне матрице $\mathbf{P}$**

у којој неће бити нултих блокова, већ ће она активно комбиновати прогнозе са апсолутно свих нивоа у систему како би се минимизовала укупна грешка.

Метод минималног трага (MinT)

Савремени приступ оптималном усклађивању развијен кроз метод минималног трага (MinT) [22]. Циљ алгоритма MinT је проналажење матрице $\mathbf{P}$ која минимизује траг (укупну дисперзију) коваријационе матрице грешака усклађених прогноза, уз услов непристрасности ($\mathbf{PS} = \mathbf{I}_m$).

Овај циљ се може формално дефинисати као следећи оптимизациони проблем:

$$\min_{\mathbf{P} | \mathbf{PS} = \mathbf{I}_m} \text{Tr}(\mathbf{SPWP}^T \mathbf{S}^T) \quad (2.14)$$

где је $\mathbf{W} = \mathbb{E}[\mathbf{e}_t \mathbf{e}_t^T]$ коваријациона матрица грешака базних прогноза ван узорка ($\mathbf{e}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t$), док $\text{Tr}(\cdot)$ означава траг матрице. Услов непристрасности ($\mathbf{PS} = \mathbf{I}_m$) представља математичко ограничење оптимизационог проблема чија је улога да спречи уношење системске грешке приликом усклађивања. Уколико базни модели генеришу непристрасне прогнозе (чије је математичко очекивање грешке једнако нули), испуњење овог услова осигурава да ће и коначне, усклађене прогнозе остати строго непристрасне за сваку појединачну серију у систему [22]. Практично посматрано, када овог ограничења не би било, алгоритам би тривијално минимизовао функцију циља (укупну дисперзију) тако што би матрицу $\mathbf{P}$ поставио на нулу, где би систем потпуно игнорисао базне прогнозе и као резултат враћао нулте (константне) вредности будући да оне немају дисперзију. Увођењем услова непристрасности, алгоритам се приморава да оптимум тражи искључиво кроз смислену линеарну комбинацију прогноза. Аналитичко решење овог оптимизационог проблема дато је једначином:

$$\mathbf{P} = (\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}^{-1}. \quad (2.15)$$

Заменом добијене матрице $\mathbf{P}$ у општу једначину, добијамо коначну формулу MinT:

$$\tilde{\mathbf{y}}_t = \mathbf{S}(\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}^{-1} \hat{\mathbf{y}}_t. \quad (2.16)$$

Апроксимације коваријационе матрице (варијанте MinT)

Суштински изазов алгоритма MinT лежи у чињеници да је стварна матрица $\mathbf{W}$ непозната и мора бити апроксимирана на основу резидуала. Резидуали представљају разлику између стварних историјских вредности и одговарајућих вредности прогнозираних моделом. Дефинисаћемо пет теоријских варијанти матрице $\mathbf{W}$, од којих је четири примењено у експерименталном делу:

- **OLS (ordinary least squares):** Ова метода претпоставља да је $\mathbf{W} \propto \mathbf{I}_n$ (матрице су пропорционалне) [9]. Ова детерминистичка метода игнорише све корелације и разлике у дисперзијама између серија. Решење се ослања искључиво на структуру хијерархије, односно груписања.
- **WLS_Struct (weighted least squares - structural):** Ова метода претпоставља да су дисперзије грешака пропорционалне броју базних серија које се сабирају у одређени чвор [9]. Матрица $\mathbf{W}$ је дијагонална: $\mathbf{W}_{struct} = \text{diag}(\mathbf{S}\mathbf{1})$, где је $\mathbf{1}$ вектор јединица. Попут OLS-а, ова метода зависи само од структуре хијерархије, односно груписања.
- **WLS_Var (weighted least squares - variance):** Емпиријска метода уведена кроз алгоритам MinT која користи израчунате дисперзије резидуала за дијагоналу, док вандијагоналне елементе (корелације) поставља на нулу [22]. Ово резултује дијагоналном матрицом $\mathbf{W}_{var} = \text{diag}(\hat{\mathbf{\Sigma}})$.
- **MinT_Sample (класична метода MinT):** Ова метода користи пуну, нерегуларизовану узорачку коваријациону матрицу грешака $\mathbf{W}_{sample} = \frac{1}{T-1} \mathbf{E}^T \mathbf{E}$ (где је $\mathbf{E}$ центрирана матрица резидуала димензија $T \times n$). Ова матрица садржи стварне емпиријске дисперзије и коваријансе без икаквих модификација. Иако је теоријски идеална, ова метода је у пракси нумерички нестабилна и често рачунарски неизводљива у великим системима. Из тог разлога, ова варијанта није уопште могла бити извршена нити обухваћена у нашем експерименталном раду.
- **MinT_Shrink (shrinkage estimator):** Савремено решење које омогућава примену пуне коваријансе у великим системима. Користи метод регуларизације који приближава нестабилну узорачку матрицу $\mathbf{W}_{sample}$

према стабилној дијагоналној матрици $\mathbf{W}_{var}$ [20]. Савремене софтверске имплементације израчунавају овај метод кроз нумеричке пречице без експлицитног креирања комплетне коваријационе матрице у меморији, чиме се избегава прекорачење расположивих рачунарских ресурса [17].

Пример израчунавања усклађених прогноза кроз 4 методе:

Да бисмо демонстрирали процес размене информација кроз хијерархију, прикажемо како се из матрица $\mathbf{W}$ добија коначна једначина за усклађену прогнозу региона А ($\hat{y}_A$) у нашем систему од 3 серије (y_T, y_A, y_B). Општа једначина за матрицу пресликавања је $\mathbf{P} = (\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}^{-1}$, при чему је

$$\text{матрица сабирања } \mathbf{S} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

1. Метода OLS:

Користи се јединична матрица $\mathbf{W}_{ols} = \mathbf{I}_3$. Када се ово уврсти, добијамо $\mathbf{P}_{ols} = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T$. Израчунавањем ове инверзије добијамо матрицу димензија 2×3 :

$$\mathbf{P}_{ols} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix} \quad (2.17)$$

Множењем првог реда ове матрице са вектором неусклађених прогноза $\hat{\mathbf{y}}_t = [\hat{y}_T, \hat{y}_A, \hat{y}_B]^T$, добијамо коначну прогнозу за регион А:

$$\tilde{y}_A = \frac{1}{3}\hat{y}_T + \frac{2}{3}\hat{y}_A - \frac{1}{3}\hat{y}_B \quad (2.18)$$

Видимо да усклађена прогноза задржава 66% сопственог сигнала, а позајмљује 33% информација са вишег нивоа, уз корекцију на основу суседног региона. Тежине су стриктно геометријске.

2. Метода WLS_Struct:

Као што је претходно дефинисано, ова метода претпоставља да је структурна дисперзија чвора пропорционална броју базних серија које се у њему сабирају. То се алгебарски израчунава множењем матрице $\mathbf{S}$ вектором јединица 1 димензија $m \times 1$ (у нашем случају 2×1 јер имамо две базне серије):

$$\mathbf{S}\mathbf{1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \quad (2.19)$$

Ово значи да укупна продаја (y_T) у себи садржи 2 базне серије (па јој је тежина 2), док су y_A и y_B саме по себи базне (па им је тежина 1). Постављањем ових вредности на дијагоналу добијамо $\mathbf{W}_{struct} = \text{diag}(2, 1, 1)$, па је њена инверзија $\mathbf{W}^{-1} = \text{diag}(\frac{1}{2}, 1, 1)$.

Множењем $(\mathbf{S}^T \mathbf{W}^{-1} \mathbf{S})$ добијамо $\begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$. Инверзијом ове матрице и множењем са $\mathbf{S}^T \mathbf{W}^{-1}$ добијамо нову матрицу пресликавања:

$$\mathbf{P}_{struct} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix} \quad (2.20)$$

Коначна прогноза сада гласи:

$$\tilde{y}_A = \frac{1}{4} \hat{y}_T + \frac{3}{4} \hat{y}_A - \frac{1}{4} \hat{y}_B \quad (2.21)$$

У поређењу са OLS-ом, сопствени сигнал базног нивоа је појачан (са 66% на 75%), док је утицај укупне продаје ($\hat{y}_T$) смањен. Разлог је то што структурна метода математички препознаје да се чвор y_T састоји од више серија, па му додељује већу дисперзију (мању поузданост у односу на директну прогнозу самог региона).

3. Метода WLS_Var:

Овде уводимо емпиријске дисперзије: $\mathbf{W}_{var} = \text{diag}(\hat{\sigma}_T^2, \hat{\sigma}_A^2, \hat{\sigma}_B^2)$. Да бисмо поједноставили запис, уведемо тежине (пондере) као реципрочну вредност дисперзије: $w_i = 1/\hat{\sigma}_i^2$. Аналитичким решавањем матричних једначина, фактор за регион А поприма следећи облик:

$$\tilde{y}_A = \frac{w_A(w_T + w_B)}{D} \hat{y}_A + \frac{w_T w_B}{D} \hat{y}_T - \frac{w_T w_B}{D} \hat{y}_B \quad (2.22)$$

где је $D = w_T w_A + w_T w_B + w_A w_B$ заједнички именилац. Ова формула осликава механизам емпиријске регуларизације. Наиме ако глобални модел на врху прави огромне грешке (варијанса $\hat{\sigma}_T^2 \rightarrow \infty$, па тежина $w_T \rightarrow 0$), фактори уз $\hat{y}_T$ и $\hat{y}_B$ постају нула, и систем одлучује да верује искључиво базном моделу ($\tilde{y}_A = \hat{y}_A$). Обрнуто, ако је глобални модел изузетно прецизан (мала дисперзија), алгоритам му даје приоритет и снажније коригује базне серије.

4. Метода MinT_Sample:

Ова метода користи пуну, нерегуларизовану коваријациону матрицу $\mathbf{W}_{sample}$. За разлику од претходних метода где је матрица била дијагонална, овде ван-дијагонални елементи (коваријансе, нпр. $\hat{\sigma}_{A,B}$) нису нуле. Због тога инверзна

матрица $\mathbf{W}_{sample}^{-1}$, као и финална матрица $\mathbf{P}_{sample} = (\mathbf{S}^T \mathbf{W}_{sample}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}_{sample}^{-1}$ постају потпуно густе матрице чији сваки елемент представља сложену функцију свих дисперзија и коваријанси у систему.

Једначина за усклађену прогнозу региона А у овом случају има облик:

$$\tilde{y}_A = \alpha \hat{y}_A + \beta \hat{y}_T + \gamma \hat{y}_B \quad (2.23)$$

где тежине α, β, γ сада зависе од корелације грешака. На пример, ако модели за регион А и регион Б конзистентно праве грешке у истом смеру (имају позитивну коваријансу $\hat{\sigma}_{A,B} > 0$), алгоритам то препознаје и користи информацију о величини грешке из региона Б да додатно „калибрише” прогнозу региона А пре него што их усклади. За наш мали систем ($n = 3$), ова метода савршено функционише. Међутим, због великог преплитања свих параметара, у великим системима ($n \gg T$) ова густа матрица постаје сингуларна и њене тежине α, β, γ се математички не могу израчунати.

5. Метода MinT_Shrink:

Како би задржала предности коваријансе из методе *MinT_Sample*, али избегла њен сингуларитет, ова метода уводи параметар скупљања $\lambda \in (0, 1)$. Регуларизована коваријациона матрица се рачуна коришћењем линеарне комбинације:

$$\mathbf{W}_{shrink} = (1 - \lambda) \mathbf{W}_{sample} + \lambda \mathbf{W}_{var} \quad (2.24)$$

Усклађена прогноза $\tilde{y}_A$ која произилази из ове матрице представља компромис између методе 3 (*WLS_Var*) и методе 4 (*MinT_Sample*). Вандијагонални елементи (коваријансе) из узорачке матрице се множе са $(1 - \lambda)$ чиме им се вештачки смањује утицај (скупљају се ка нули), док се стабилне дисперзије на дијагонали појачавају. Алгоритам оптимално бира λ тако да систем научи унакрсне корелације између серија (нпр. између региона), али да матрица истовремено остане довољно дијагонална да би се гарантовала њена успешна инверзија у оптимизационом процесу.

Рачунарски изазови, нумеричка стабилност и проблем ненегативности

Иако алгебарски оквир усклађивања MinT нуди елегантно теоријско решење, његова примена на реалне, високодимензионалне пословне системе открива низ критичних математичких и инжењерских ограничења. Ови изазови

диктирају применљивост алгоритама у пракси и обликују методологију овог рада.

Проблем сингуларитета и прекорачења меморије

Први суштински изазов јавља се при покушају процене пуне коваријационе матрице $\mathbf{W}_{sample}$ (класични MinT). Када је број серија n значајно већи од броја историјских тренутака T (што је правило у ланцима снабдевања), ова $n \times n$ матрица је непотпуног ранга (сингуларна) [22]. Строго математички посматрано, ранг ове матрице је ограничен на $T - 1$, будући да центрирање резидуала приликом рачунања коваријансе уводи линеарну зависност и одузима један степен слободе. Пошто је $T \ll n$, њена инверзија математички не постоји.

Поред математичке немогућности, постоји и физичко ограничење рачунарских система. Експлицитна алокација густе коваријационе матрице димензија $n \times n$ захтева гигантске количине радне меморије (често десетине гигабајта). Ово моментално доводи до грешке прекорачења меморије и пада система. За разлику од ње, детерминистичке и дијагоналне методе (OLS, WLS) немају овај проблем јер се у меморији чувају само као једнодимензионални низови (дијагонале) и задржавају пун ранг. Метод *MinT_Shrink* успешно заобилази проблем прекорачења меморије јер се, захваљујући регуларизацији и оптимизованим имплементацијама, комплетан поступак извршава директно над „уском” матрицом резидуала $\mathbf{E}$ димензија $T \times n$, са минималним меморијским захтевима.

Проблем ненегативности и декомпозиције Чолеског

Други, једнако важан изазов произилази из чињенице да основна формулација алгорита MinT представља неограничену линеарну оптимизацију. Примена овакве трансформације често генерише негативне усклађене прогнозе ($\tilde{y}_{i,t} < 0$), што је у контексту пословне тражње физички немогуће и негативно утиче на метрике перформанси.

Теоријски, овај проблем се решава увођењем формалних ограничења ненегативности ($\tilde{\mathbf{y}}_t \geq \mathbf{0}$), што аналитичко решење једначине MinT претвара у проблем квадратног програмирања (енг. *quadratic programming* - *QP*). Међутим, примена QP решавача (енг. *solvers*) захтева итеративну инверзију и факторизацију матрица, најчешће коришћењем декомпозиције Чолеског ($\mathbf{Q} = \mathbf{L}\mathbf{L}^T$, где је $\mathbf{Q}$ матрица квадратног проблема, тј. хесијан). Да би се ова

декомпозиција извршила, матрица $\mathbf{Q}$ мора бити позитивно семидефинитна.

У пракси, висока испрекиданост серија генерише изузетно мале варијансе резидуала. Како у овом конкретном оптимизационом проблему хесијан поприма облик који директно зависи од инверзне матрице коваријансе ($\mathbf{Q} \propto \mathbf{S}^T \mathbf{W}^{-1} \mathbf{S}$), ове вредности блиске нули након инверзије теже бесконачности. Иако је хесијан теоријски позитивно семидефинитан, услед комбиновања оваквих огромних и јако малих вредности (код стабилних серија на вишим агрегационим нивоима) у математичким операцијама, ограничена рачунарска прецизност доводи до грешака у заокруживању, због којих матрица може добити негативне сопствене вредности [16]. Последица је да није могуће извршити декомпозицију Чолеског, а самим тим ни оптимизацију. За разлику од њих, детерминистичке методе (OLS, WLS_Struct) не зависе од варијанси већ од стабилне структуре хијерархије, те успешно конвергирају.

Компромис кроз хеуристичко одсецање негативних вредности

Као практична алтернатива формалној оптимизацији код рачунарски нестабилних емпиријских матрица, у овом раду се примењује стратегија хеуристичког одсецања негативних вредности на нулу. Овде се, након примене стандардне неограничене једначине MinT, свака вредност $\tilde{y}_{i,t} < 0$ експлицитно пресликава у 0. Иако овај захват нарушава савршену кохерентност збирова унутар матрице $\mathbf{S}$, он отвара питање: да ли се вредност хијерархијског усклађивања налази искључиво у савршеној математичкој кохерентности, или у чињеници да метод (непосредно пре одсецања негативних вредности) успешно врши трансфер стабилних информација из виших нивоа ка базном нивоу.

Глава 3

Дизајн истраживања и методологија

Ово поглавље описује експериментални оквир коришћен за испитивање утицаја хијерархијског усклађивања на тачност предвиђања временских серија. Методологија је пројектована тако да истовремено задовољи два захтева: да остане математички доследна формалном оквиру усклађивања дефинисаном у претходном поглављу, а са друге стране, да буде изводљива над реалним, високодимензионим подацима са испрекиданом потражњом из пословног окружења. Због тога се посебна пажња посвећује конструкцији хијерархије, стратегији валидације, избору предиктивних модела, начину процене резидуала и постпроцесирању прогноза које нарушавају природно ограничење ненегативности. Практична реализација свих наведених корака софтверски је имплементирана у програмском језику *Python*. Хардверско окружење у којем су експерименти извршавани састоји се од 32 GB оперативне меморије (RAM), процесора Intel Core i7-11850H (2.50 GHz) и графичке картице NVIDIA RTX A3000 са 6 GB видео меморије (VRAM).

3.1 Карактеристике и припрема скупа података

Анализа је спроведена над приватним подацима компаније Bosch из домена тржишта резервних делова за аутомобилску индустрију. Посматрани регион је Северна Америка, а временска фреквенција оригиналних података

је месечна. Скуп података покрива период од 2020. до 2025. године (5 година историјских података) уз истовремено задржавање типичних ограничења реалних пословних система: променљив интензитет потражње и значајан број ретких производа.

Базни ниво хијерархије дефинисан је комбинацијом три идентификационе колоне: $ts_id = (\text{производ}, \text{купац}, \text{локација})$.

Другим речима, свака базна временска серија представља месечну потражњу за једну комбинацију производа, купца и локације. У имплементацији се ове три колоне спајају у јединствени идентификатор серије, означен са ts_id . Након филтрирања задржано је приближно 20.000 конзистентних базних серија. Под конзистентношћу се подразумева да је свака серија довољно дуга, да се може једнозначно пресликати у све посматране агрегационе нивое и да не нарушава алгебарску структуру матрице сабирања $\mathbf{S}$ у посматраном временском периоду (2020. - 2025.).

Нека је $y_{i,t}$ вредност потражње за базну серију i у месецу t . За сваки временски тренутак формира се вектор базних вредности

$$\mathbf{b}_t = \begin{bmatrix} y_{1,t} & y_{2,t} & \cdots & y_{m,t} \end{bmatrix}^T, \quad (3.1)$$

где је m број базних серија. Све агрегиране серије добијају се детерминистички применом матрице сабирања:

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t. \quad (3.2)$$

Ова формулација обезбеђује да су историјске вредности потпуно кохерентне пре почетка предвиђања. На тај начин се сва некохерентност која се јавља у експерименту може приписати независно генерисаним прогнозама и накнадном поступку усклађивања, а не грешкама у припреми података.

Дизајн структурне хијерархије

Хијерархијска структура у раду није ограничена на једно стабло, већ комбинује производну хијерархију са више независних групационих погледа. Базни ниво је дефинисан комбинацијом **производ**, **купац** и **локација**, док се виши нивои добијају агрегацијом по појединачним пословним атрибутима.

Разматране су следеће структурне хијерархије:

- **купац**: независна агрегација по купцу са 127 серија. Све базне серије које припадају истом купцу сабирају се у једну серију на нивоу купца.

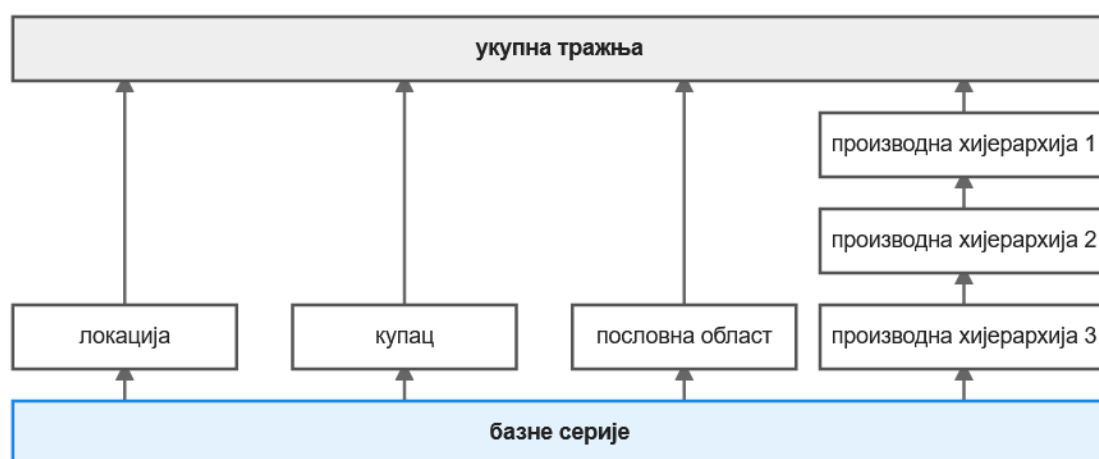
- **локација:** независна агрегација по локацији, са 10 серија. Овај ниво групише потражњу према географској или логистичкој локацији.
- **пословна област:** независна агрегација по пословним областима, са 16 серија. Овај поглед представља пословну класификацију.
- **производна хијерархија 1 → производна хијерархија 2 → производна хијерархија 3:** производна хијерархија у три нивоа. За разлику од претходних групација, ови нивои имају природан угнежђени однос: производна хијерархија 3 представља најдетаљнији производни ниво у овој групи, са 686 серија; производна хијерархија 2 је његов родитељски ниво, са 396 серија; док је производна хијерархија 1 најагрегиранији производни ниво, са 119 серија.
- **укупно:** укупни ниво система, који сабира све базне серије у једну временску серију.

Природно је прво посматрати најширу доступну структурну поставку, јер она садржи све пословне погледе кроз које се потражња може агрегирати: купца, локацију, пословну област, производну хијерархију и укупни ниво. Међутим, укључивање свих структурних нивоа истовремено повећава број агрегираних серија и број веза кроз које се информација размењује током усклађивања. Због тога је, поред пуне поставке, оправдано размотрити и ужи подскуп хијерархија.

Ужа поставка обухвата нивое *локација*, *пословна област*, *производна хијерархија 1* и *укупно*. Овај избор задржава различите типове агрегационе информације: географско-логистички поглед, категорички пословни поглед, најагрегиранији производни поглед и укупни системски ниво. Истовремено се изостављају нивои са већим бројем серија, као што су *купац*, *производна хијерархија 2* и *производна хијерархија 3*, чиме се добија једноставнија структура за проверу осетљивости усклађивања на број укључених нивоа и густину њихових односа. На тај начин се у методолошки дизајн уводе две елементарне поставке: пуна, која максимално користи доступну хијерархијску информацију, и редукована, која проверава понашање усклађивања када је размена информација ограничена на мањи број агрегираних погледа.

Алгебарски, сви ови нивои могу се представити истом матрицом сабирања **S**, али се њихова семантика разликује (слика 3.1). Нивои *купац*, *локација* и

пословна област представљају независне групације над истим базним скупом серија. То значи да једна базна серија истовремено припада једном купцу, једној локацији и једној пословној области, али ови атрибути не формирају међусобно угнежђено стабло. Насупрот томе, производни нивои *производна хијерархија 1*, *производна хијерархија 2* и *производна хијерархија 3* третирају се као угнежђена производна хијерархија.



Слика 3.1: Приказ хијерархијске структуре. Независни атрибути (локација, купац, пословна област) директно агрегирају базне серије ка укупној тражњи, док производни атрибути формирају секвенцијално угнежђено стабло.

Због тога се целокупна структура може посматрати као комбинација једне стриктне производне хијерархије и више независних груписаних хијерархија. Ако је $\mathbf{b}_t$ вектор базних серија, сваки структурни поглед дефинише сопствени блок редова у матрици $\mathbf{S}$. На пример, ред који одговара једном купцу има јединице на позицијама свих базних серија које припадају том купцу, док су остали елементи нуле. Аналогно томе, ред за једну локацију, пословну област или производни ниво сабира одговарајући подскуп базних серија.

Временска агрегација и режими усклађивања

Поред структурне агрегације, у анализу је укључена и временска агрегација. Оригиначне месечне серије агрегирају се на кварталну фреквенцију, означену са $\mathbf{QS}$. Пошто је хоризонт месечног предвиђања $H = 6$, одговарајући квартални хоризонт је $H_q = 2$. За сваку серију квартална вредност добија се

сабирањем три узастопне месечне вредности:

$$y_{i,r}^{(q)} = y_{i,3r-2}^{(m)} + y_{i,3r-1}^{(m)} + y_{i,3r}^{(m)}, \quad r = 1, \dots, H_q. \quad (3.3)$$

Временска агрегација се примењује на базни ниво и на сваки структурни ниво посебно. То значи да се, поред месечних серија на свим нивоима посматрају и њихове кварталне верзије. Међутим, временско усклађивање није спроведено као једно заједничко структурно-временско усклађивање над свим нивоима истовремено. Уместо тога, сваки попречни пресек хијерархије усклађује се одвојено са својом кварталном верзијом. За сваки ниво формира се засебан временски систем у ком вектор прогноза садржи кварталне и месечне прогнозе исте серије:

$$\hat{\mathbf{y}}_i = \begin{bmatrix} \hat{\mathbf{y}}_i^{(q)} \\ \hat{\mathbf{y}}_i^{(m)} \end{bmatrix}. \quad (3.4)$$

Матрица сабирања у овом случају описује само временску везу између месечних и кварталних корака. За хоризонт од шест месеци и два квартала она има облик:

$$\mathbf{S}_{\text{temp}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (3.5)$$

Прва два реда представљају кварталне агрегате, док последњих шест редова представљају месечне вредности. За сваки појединачни ниво и сваку серију временско усклађивање се зато своди на пројекцију независних месечних и кварталних прогноза у простор дефинисан овом матрицом.

Потпуно заједничко структурно-временско усклађивање разматрано је као проширење ове поставке. У том случају би једна матрица сабирања $\mathbf{S}_{\text{joint}}$ истовремено садржала структурне односе између базног нивоа и агрегата, као и временске односе између месечних и кварталних прогноза. Формално, колоне такве матрице одговарале би базним месечним величинама, док би редови обухватили:

- месечне вредности на базном нивоу;

- месечне вредности на свим структурним агрегатима;
- кварталне вредности на базном нивоу;
- кварталне вредности на свим структурним агрегатима.

Међутим, ова формулација је у посматраном систему довела до проблема прекорачења радне меморије. Разлог је у томе што се број чворова у заједничком систему не повећава линеарно само бројем серија, већ и бројем временских корака и бројем структурних погледа. На пример, ако је број базних серија приближно $m = 20.000$, а месечни хоризонт $H = 6$, број колона заједничке матрице сабирања одговара броју базних месечних величина:

$$mH = 20.000 \cdot 6 = 120.000. \quad (3.6)$$

Број редова је већи, јер обухвата све величине које се изражавају преко базних месечних вредности: базне месечне прогнозе, базне кварталне прогнозе, структурне месечне агрегате и структурне кварталне агрегате. Приближно, овај број износи 170.000. Због тога је заједничка матрица сабирања димензија:

$$\mathbf{S}_{\text{joint}} \in \mathbb{R}^{170.000 \times 120.000}. \quad (3.7)$$

Ова процена не укључује додатне матрице које настају током пројекције MinT, факторизације или процене коваријансе, па би стварни меморијски захтев био још већи. Због тога је заједничко структурно-временско усклађивање искључено из финалног експерименталног извршавања. Уместо њега, примењена су два рачунарска изводљива режима: структурно усклађивање на истој фреквенцији и временско усклађивање појединачно за сваки попречни пресек. Овај избор задржава могућност анализе оба механизма преноса информација, али избегава експлицитну конструкцију огромне заједничке матрице која није изводљива на расположивим ресурсима.

3.2 Поставка и процес спровођења експеримента

Стратегија тренирања, валидације и оптимизације хиперпараметара

Полазна тачка експерименталне поставке је хронолошка подела података на тренинг и тест период. Сви модели се обучавају искључиво на историјским опажањима које претходе тест периоду, док је последњих шест месеци посматраног периода издвојено за коначну евалуацију. На тај начин се симулира реалан сценарио прогнозирања, у ком модел у тренутку обучавања нема приступ будућим вредностима. Хоризонт предвиђања је зато $H = 6$.

За оптимизацију хиперпараметара, тренинг скуп се додатно дели на интерни скуп за обучавање и валидациони прозор. Валидациони прозор дефинисан је као последњих шест месеци тренинг скупа, чиме се симулира исти тип задатка као у коначном тестирању, али без коришћења информација из тест периода. Оптимизација хиперпараметара спроводи се као бајесовска оптимизација над унапред дефинисаним простором претраге. За сваку комбинацију модела и нивоа хијерархије покреће се ограничен број покушаја, при чему је у експерименталној конфигурацији број покушаја постављен на 10.

Због великог броја базних серија и високих трошкова обучавања, оптимизација хиперпараметара се на базном нивоу спроводи над подузорком од 10% серија. На агрегираним нивоима користи се пун скуп серија, јер је број серија на тим нивоима значајно мањи и рачунарски изводљив за директну оптимизацију. Овај приступ представља компромис између репрезентативности и рачунарске захтевности.

Критеријум оптимизације је WMAPE (енг. weighted mean absolute percentage error) на валидационом прозору:

$$\text{WMAPE} = \frac{\sum_{i,t} |y_{i,t} - \hat{y}_{i,t}|}{\sum_{i,t} |y_{i,t}| + \varepsilon}, \quad (3.8)$$

где је ε мала позитивна константа која спречава дељење нулом код серија са изузетно ретком потражњом.

Након избора најбоље комбинације хиперпараметара, модел се поново обучава над целокупним тренинг скупом одговарајућег нивоа хијерархије, односно над свим опажањима које претходе тест периоду. Овим кораком се избе-

гава губитак информација који би настао када би финални модел остао обучен искључиво на редукованом скупу података издвојеном за оптимизацију хиперпараметара. Дакле, валидациони прозор служи за избор хиперпараметара, али се након тог избора подаци из тог прозора враћају у тренинг скуп за финално обучавање модела.

Код модела дубоког учења задржава се посебан део тренинг скупа за праћење вредности функције губитка током обучавања. Овај интерни валидациони скуп дефинисан је да буде исте дужине као и хоризонт предвиђања, што у овом експерименту износи $H = 6$ временских периода. Овај део не представља коначни тест скуп, већ интерни валидациони сигнал који омогућава праћење стабилности обучавања, рано заустављање и контролу прекомерног прилагођавања. На тај начин се код неуронских модела раздвајају три улоге података: тест период за коначну евалуацију, валидациони скуп за избор хиперпараметара и интерни валидациони сигнал током самог обучавања.

Тако добијени финални модел затим генерише прогнозу за наредних $H = 6$ месеци:

$$\hat{\mathbf{y}}_{T+1:T+H}. \quad (3.9)$$

Ове прогнозе се, у зависности од експерименталног режима, или директно евалуирају као неусклађене прогнозе или прослеђују поступку хијерархијског усклађивања.

Предиктивни модели

Статистички модели

Прву групу сачињавају класични статистички модели. Њихова основна карактеристика је локални приступ, што значи да се параметри модела оптимизују потпуно независно за сваку појединачну временску серију:

- **AutoARIMA:** Коришћењем алгорита из библиотеке `pmdarima`, за сваку серију се аутоматски идентификује оптимална структура модела на основу Акаикеовог информационог критеријума (енг. *Akaike information criterion* - *AIC*). Код овог модела уведен је резервни механизам за случајеве нумеричких отказа, попут немогућности конвергенције оптимизационог алгорита код изузетно испрекиданих серија. Уколико алгоритам откаже, за ту серију се локално додељује последња

посматрана вредност ($\hat{y}_{i,T+h} = y_{i,T}$), чиме се спречава прекид рада и осигурава потпуна покривеност система.

- **Holt-Winters (Експоненцијално равнање):** Имплементиран преко класе `ExponentialSmoothing` из библиотеке `statsmodels`. С обзиром на природу података и присуство нултих вредности у испрекиданим серијама, у експерименту су стриктно коришћени адитивни тренд и адитивна сезоналност, будући да мултипликативне компоненте нису математички дефинисане над серијама које садрже нуле.
- **Prophet:** Имплементиран је коришћењем званичне `prophet` библиотеке. Овај алгоритам приступа проблему предвиђања кроз декомпозицију на тренд и сезоналност коришћењем генерализованих адитивних модела (GAM) [21]. Суштински, Prophet моделује временску серију као збир три главне компоненте: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$. Компонента $g(t)$ представља тренд (најчешће део по део линеарну или логистичку криву) који се аутоматски прилагођава структурним променама у подацима. Компонента $s(t)$ моделује периодичне сезонске обрасце коришћењем Фурије-ових редова који не дозвољавају велики утицај одударајућих вредности, док $h(t)$ обухвата ефекте празника и специфичних пословних догађаја. Prophet проблем предвиђања своди на прилагођавање криве, а не на строгу ауторегресију, због чега је отпоран на недостајуће вредности јер функцију прилагођава искључиво доступним опажањима.

Иако се статистички модели обучавају локално, њихова оптимизација (НРО) спроводи се глобално. Како би се спречило преприлагођавање (енг. *overfitting*) локалном шуму, одређени хиперпараметри се фиксирају и заједнички оптимизују за цео ниво хијерархије.

Предност ових модела је то што су интерпретабилни, међутим, њихов фундаментални недостатак јесте то што се свака серија третира независно, па ови модели не могу да позајмљују статистичку снагу из сродних серија током процеса обучавања (енг. *cross-learning*).

Класични модели машинског учења

Другу групу чине глобални регресиони модели засновани на ансамблима стабала одлучивања:

- **Random Forest:** Имплементиран преко класе `RandomForestRegressor` из библиотеке `scikit-learn`. Заснива се на техници просте агрегације независних стабала одлучивања (енг. *bagging*) ради смањења дисперзије [4].
- **LightGBM:** Имплементиран преко класе `LGBMRegressor` из библиотеке `lightgbm`. Заснива се на алгоритму градијентног појачавања (енг. *gradient boosting*) који секвенцијално гради ансамбл са циљем минимизације резидуалне грешке [11].

За разлику од статистичких модела који директно раде са временским нивоима, модели засновани на стаблима захтевају трансформацију временских серија у табеларни формат. За оба модела користи се заједнички механизам конструкције карактеристика (енг. *feature engineering*). За серију i и тренутак t , улазни вектор $\mathbf{x}_{i,t}$ садржи блок историјских вредности (енг. *lags*) и аутоматски екстраховане календарске карактеристике:

$$\mathbf{x}_{i,t} = \left[y_{i,t-p}, \dots, y_{i,t-1}, \text{month}(t), \text{quarter}(t), \text{year}(t) \right]^T \quad (3.10)$$

где p представља максимални број историјских вредности који је укључен у модел као хиперпараметар.

Ови модели се обучавају глобално, што значи да један јединствени модел користи улазне узорке ($\mathbf{x}_{i,t}$) из свих 20.000 серија истовремено. На тај начин, стабла одлучивања обједињују информације из свих серија и способна су да препознају заједничке нелинеарне обрасце тражње на нивоу целог система.

У погледу генерисања прогноза кроз време, у експерименту се испитују две алтернативне стратегије предвиђања вишеструких корака унапред:

- **Рекурзивна стратегија:** Обучава се један једнокорачни модел (модел предвиђа само $t + 1$). За хоризонте веће од 1, модел користи сопствене претходно предвиђене вредности као историјске вредности за следеће кораке.
- **Директна стратегија:** Обучава се потпуно засебан модел за сваки корак хоризонта h . Ова стратегија је рачунарски знатно захтевнија, али избегава акумулацију грешке до које долази услед рекурзивног враћања сопствених (потенцијално нетачних) прогноза.

Модели дубоког учења

Трећу групу чине напредни глобални модели засновани на архитектурама дубоког учења (енг. *deep learning*):

- **DeepAR:** Имплементиран преко класе `DeepAREstimator` из библиотеке `GluonTS`. Овај алгоритам представља ауторегресивну архитектуру засновану на рекурентним неуронским мрежама (најчешће са LSTM ћелијама) [19]. Уместо да директно предвиђа будућу тачкасту вредност, мрежа у сваком временском кораку t узима скривено стање из претходног корака $\mathbf{h}_{i,t-1}$, претходно опажање $y_{i,t-1}$ и друге променљиве $\mathbf{x}_{i,t}$ (попут календарских одлика) како би проценила параметре унапред дефинисане вероватносне расподеле (нпр. средњу вредност μ и дисперзију σ^2 код Гаусове расподеле). Предност овог приступа је што модел учи заједничке тежине мреже из свих серија, док интерна локална скалирања омогућавају прилагођавање специфичним магнитудама тражње сваког производа. Како би се излаз овог вероватносног модела могао интегрисати са методама хијерархијског усклађивања (MinT), из генерисане предиктивне расподеле се за потребе даље оптимизације екстрахује искључиво њено математичко очекивање (средња вредност).
- **N-HITS (neural hierarchical interpolation for time series):** Имплементиран преко `NeuralForecast` библиотеке из Nixtla екосистема. Овај модел представља најсавременију блоковску неуронску архитектуру која решава проблеме комплексности традиционалних рекурентних мрежа увођењем вишеструких стопа узорковања [5]. Архитектура се састоји од низа узастопних слојева (блокова), од којих сваки обрађује сигнал на различитој фреквенцији. Коришћењем операција агрегације максимумом (енг. *max-pooling*), почетни блокови екстрахују дугорочне, нискофреквентне трендове из улазног сигнала. Резидуална грешка се прослеђује каснијим блоковима који се фокусирају на све финије, високофреквентне локалне флукуације. Овако екстраховани сигнали са различитих скала се затим спајају помоћу интерполације (формирајући коначну прогнозу). Осим што пружа високу интерпретабилност кроз декомпозицију сигнала, N-HITS генерише вишекорачне прогнозе директно (сви кораци хоризонта H се генеришу истовремено применом хијерархијске интерполације екстрахованих сигнала), чиме се у потпуности избе-

гава акумулација грешке карактеристична за рекурзивне ауторегресивне моделе.

Као и у случају стабала одлучивања, обе неуронске архитектуре се обучавају глобално, над свим серијама у систему истовремено.

Вишенаменски претренирани модели

Четврту и најсавременију групу предиктивних алгоритама у овом експерименту чине велики, вишенаменски претренирани модели:

- **Chronos (Chronos-2)**: Овај алгоритам припада најновијој генерацији вишенаменских претренираних модела развијених од стране компаније Amazon. Chronos приступа предвиђању временских серија кроз призму процесирања природног језика (NLP), користећи класичну трансформер архитектуру [1]. Нумеричке вредности временске серије се прво дискретизују у специфичне токене, чиме се проблем своди на класичан проблем класификације следећег токена у језичком моделу.
- **TimesFM (time series foundation model)**: Модел развијен од стране компаније Google. За разлику од Chronos-а који вредности претвара у токене, TimesFM користи архитектуру засновану искључиво на декодеру (слично као GPT модели), уз примену технике дељења сигнала на делове (енг. *patching*) [7]. Ова техника му омогућава да директно обрађује сирове непрекидне вредности кроз блокове података, чиме задржава високу ефикасност и брзину при раду са серијама различитих фреквенција.

За разлику од статистичких модела и модела машинског учења, фаза обучавања код вишенаменских претренираних модела не постоји у традиционалном смислу. Нема ажурирања параметара мреже нити прилагођавања тежина на специфичном скупу података. Уместо тога, фаза припреме подразумева само учитавање претходно обучених тежина модела и формирање одговарајућих контекстних прозора (енг. *context windows*) који садрже историјске вредности серије. Систем затим на основу тог прозора директно генерише будуће вредности.

Постпроцесирање прогноза: одсецање негативних вредности

Пошто посматрана величина представља потражњу, природно ограничење система је ненегативност ($y_{i,t} \geq 0$). Међутим, већина примењених базних модела, као и сами линеарни оператори усклађивања, не гарантују да ће излазне прогнозе остати ненегативне.

Формално увођење ограничења ненегативности претвара аналитичко решење MinT у проблем квадратног програмирања (QP). Као што је раније дискутовано, на реалним подацима са великим бројем испрекиданих серија, емпиријске матрице тежина ($\mathbf{W}_{var}$ и $\mathbf{W}_{shrink}$) често постају лоше условљене, што током факторизације у QP решавачима узрокује нумеричку нестабилност и отказ система.

Због тога је експериментални поступак хијерархијског усклађивања подељен у два правца:

1. За рачунарски стабилне детерминистичке методе (OLS и WLS_Struct), усклађивање је спроведено на два паралелна начина: коришћењем формалне QP оптимизације са ограничењима, као и коришћењем неограниченог усклађивања са накнадним хеуристичким одсецањем негативних вредности.
2. За све методе (укључујући нестабилне емпиријске методе WLS_Var и $MinT_Shrink$), спроведено је неограничено усклађивање након којег је примењен хеуристички постпроцесор који све негативне вредности пресликава у нулу:

$$\tilde{y}_{i,t}^+ = \max(\tilde{y}_{i,t}, 0). \quad (3.11)$$

Оваква експериментална поставка омогућава да се у наредном поглављу егзактно упореди утицај строге математичке кохерентности (QP ограничења) у односу на једноставну пост-хок хеуристику (одсецање негативних вредности на нулу), као и њихов утицај на крајњу прецизност.

Критеријуми и метрике за евалуацију

Перформансе предиктивног система процењују се поређењем стварних историјских опажања $y_{i,t}$ са свим варијантама генерисаних прогноза. При-

ликом избора евалуационих метрика, посебна пажња посвећена је њиховој робусности, тј. способности да се примене на серије драстично различитих размера и да остану нумерички стабилне у условима високе испрекиданости и присуства нултих вредности у имениоцу.

Основна метрика коришћена током интерне оптимизације хиперпараметара (НРО), као и за примарно рангирање модела, јесте **WMAPE** (пондерисана апсолутна процентуална грешка):

$$\text{WMAPE} = \frac{\sum_{i,t} |y_{i,t} - \hat{y}_{i,t}|}{\sum_{i,t} |y_{i,t}| + \varepsilon}. \quad (3.12)$$

За свеобухватну финалну евалуацију на тест скупу, у раду се прати шири скуп метрика дефинисан у оквиру евалуационог модула. Нека је N укупан број тачака за које се рачуна прогноза. Метрике су дефинисане на следећи начин:

- **nRMSE (нормализовани корен средње квадратне грешке):**

Мери стандардну девијацију резидуала, нормализовану средњом вредношћу стварних опажања како би се омогућило поређење серија различитог обима потражње:

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{N} \sum_{i,t} (y_{i,t} - \hat{y}_{i,t})^2}}{\frac{1}{N} \sum_{i,t} y_{i,t} + \varepsilon} \quad (3.13)$$

- **sMAPE (симетрична апсолутна процентуална грешка):** Решава проблем асиметричности MAPE метрике (која пенализује позитивне грешке више него негативне) тако што у имениоцу укључује и прогнозу:

$$\text{sMAPE} = \frac{1}{N} \sum_{i,t} \frac{2|y_{i,t} - \hat{y}_{i,t}|}{|y_{i,t}| + |\hat{y}_{i,t}| + \varepsilon} \quad (3.14)$$

- **BIAS_PCT (процентуална пристрасност):** Показује усмереност грешке, односно да ли модел на нивоу система конзистентно прецењује (позитиван bias) или потцењује (негативан bias) стварну потражњу:

$$\text{BIAS_PCT} = \frac{\sum_{i,t} (\hat{y}_{i,t} - y_{i,t})}{\sum_{i,t} y_{i,t} + \varepsilon} \quad (3.15)$$

Како би се детаљно испитао утицај сваког појединачног корака у експерименталном процесу, све наведене метрике се рачунају и упоређују у четири различите фазе:

1. На сировим (неусклађеним) базним прогнозама ($\hat{\mathbf{y}}$).
2. На прогнозама усклађеним класичним неограниченим алгоритмом ($\tilde{\mathbf{y}}$).
3. На прогнозама усклађеним уз формална QP ограничења ненегативности (за рачунарски стабилне методе).
4. На усклађеним прогнозама над којима је примењено хеуристичко одсецање негативних вредности на нулу ($\tilde{\mathbf{y}}^+$).

Важно је прецизирати начин на који се све наведене метрике израчунавају и агрегирају у коначне вредности које се извештавају у резултатима. У овом раду примењен је приступ макро-усредњавања (енг. *macro-averaging*). Евалуациони модул прво израчунава све метрике независно за сваку појединачну серију i кроз задати хоризонт предвиђања. На пример, WMAPE за једну специфичну серију рачуна се као:

$$\text{WMAPE}_i = \frac{\sum_{t=T+1}^{T+H} |y_{i,t} - \hat{y}_{i,t}|}{\sum_{t=T+1}^{T+H} y_{i,t} + \varepsilon}. \quad (3.16)$$

Након израчунавања метрика на нивоу појединачних серија, коначна оцена за одређени агрегациони ниво (за дату комбинацију модела и фазе евалуације) добија се израчунавањем аритметичке средине (просека) тих вредности, уз игнорисање недефинисаних вредности:

$$\text{WMAPE}_{\text{overall}} = \frac{1}{N} \sum_{i=1}^N \text{WMAPE}_i \quad (3.17)$$

Овакав дизајн евалуације има оправдање у домену ланаца снабдевања, где за разлику од приступа глобалног сумирања (енг. *micro-averaging*) који интерно фаворизује серије са великим обимом продаје (које нису нужно битније), макро-усредњавање даје једнаку тежину свим серијама у систему.

Глава 4

Резултати и дискусија

У овом поглављу приказани су резултати експерименталне евалуације. Анализа је подељена у неколико логичких целина како би се систематски одговорило на истраживачка питања дефинисана у уводу. Прво су анализирани перформансе сирових, неусклађених модела, чиме је постављена основа за мерење ефеката усклађивања. Затим је спроведена свеобухватна анализа структурног и темпоралног усклађивања, са посебним фокусом на утицај доменских ограничења (хеуристичког одсецања негативних вредности и формалног квадратног програмирања). На крају, квантификован је губитак математичке кохерентности, након чега су формулисане најбоље праксе за индустријску примену.

4.1 Анализа перформанси неусклађених предвиђања

Пре евалуације ефеката хијерархијског усклађивања, неопходно је установити апсолутне предиктивне перформансе основних модела. Резултати приказани у овом одељку служе као контролна основа у односу на коју се рачунају сва каснија маргинална побољшања или деградације. Евалуација је спроведена на најнижем, базном нивоу (приближно 20.000 временских серија).

Као што је приказано у табели 4.1, резултати евалуације базних модела указују на изузетно велику разлику у тачности, при чему је најлошији модел готово осам пута непрецизнији од најбољег. Могу се издвојити два кључна

Табела 4.1: Упоредни приказ перформанси неусклађених базних модела (сортирано по WMAPE метрици).

Базни модел	WMAPE	nRMSE	sMAPE	BIAS_PCT (%)
<i>DeepAR</i>	110.2	142.4	108.0	-4.2
<i>N-HITS</i>	113.4	143.2	116.3	6.5
<i>TimesFM</i>	118.6	147.8	112.0	15.4
<i>Chronos</i>	126.3	157.9	118.2	20.9
<i>Holt-Winters</i>	189.0	235.0	103.6	104.8
<i>Prophet</i>	190.8	228.2	103.9	103.8
<i>ARIMA</i>	192.4	217.6	106.5	120.4
<i>Random Forest</i>	301.6	318.6	106.8	249.9
<i>Random Forest (direct)</i>	304.1	322.9	106.3	256.1
<i>LightGBM</i>	865.4	878.5	120.3	832.7
<i>LightGBM (direct)</i>	894.0	908.0	120.5	863.8

обрасца:

- Доминација дубоког учења и вишенаменских претренираних модела:** Најуспешнији предиктивни алгоритми на сировим подацима припадају фамилији неуронских мрежа. Модели *DeepAR* и *N-HITS* остварили су најбоље резултате. Изузетно је значајан и резултат вишенаменских претренираних модела, *TimesFM* и *Chronos*, који су остварили врхунске перформансе у режиму без икаквог дообучавања на специфичним подацима компаније. Овај резултат недвосмислено потврђује њихову снажну способност генерализације образаца чак и на високо испрекиданим серијама.
- Изазови примене модела заснованих на стаблима:** Модели класичног машинског учења (*LightGBM* и *Random Forest*) остварили су знатно лошије резултате, уз генерисање велике позитивне пристрасности. Ово не указује да су сами алгоритми инхерентно лоши, већ је директна последица њихове стандардне функције циља MSE (енг. mean squared error). С обзиром на то да MSE оптимизује очекивану средњу вредност, модел на испрекиданим серијама често не успева да предвиди нулу у периодима мировања, већ генерише константну позитивну вредност чије акумулирање кроз периоде без потражње може довести до позитивне пристрасности. Овакво понашање модела је условљено

уделом нултих вредности, и у пракси би се могло ублажити одабиром адекватније функције циља. Поред тога, за ове моделе коришћен је релативно скроман скуп улазних карактеристика. Они намерно нису додатно прилагођавани нити обогаћивани новим подацима, будући да је један од циљева експеримента био управо да се анализира понашање и робусност алгоритама за хијерархијско усклађивање у ситуацијама када су на улазу присутне слабије и високо пристрасне базне прогнозе.

Овакав распоред снага представља идеалан полигон за тестирање алгоритама MinT, јер систем садржи како добре, робусне моделе, тако и моделе са лошијим перформансама.

4.2 Упоредна анализа метода за усклађивање

У овом одељку анализира се утицај различитих метода усклађивања на прецизност базних прогноза. Резултати су подељени у зависности од тога да ли се усклађивање вршило кроз структурну или временску димензију.

Како би се омогућио транспарентан увид у промене свих метрика, резултати су визуелизовани кроз серију топлотних мапа које су приказане засебно за сваку димензију. Свака димензија садржи по две слике: прва приказује стабилне методе (*OLS* и *WLS_Struct*) за које је било могуће извршити сва три приступа (неограничено *reg*, хеуристичко одсецање негативних вредности *clip*, и формална оптимизација *qp*); друга приказује емпиријске методе (*MinT_Shrink* и *WLS_Var*) где QP приступ није приказан услед нумеричких сломова.

На овим приказима, редови означавају 11 базних предиктивних модела, док колоне представљају примењени приступ усклађивању. Свака ћелија топлотне мапе садржи две вредности: горњи број представља апсолутну разлику у односу на неусклађени базни модел, док доњи број у загради представља процентуалну промену. Негативне вредности (зелена боја) указују побољшање метрике након усклађивања, док позитивне вредности (црвена боја) означавају деградацију.

ГЛАВА 4. РЕЗУЛТАТИ И ДИСКУСИЈА

CSM 3-Approach Comparison: reg vs clip vs qp (Full Hierarchy, Base Level)													
Variants with QP available green = improvement, red = degradation													
OLS		WMAPE			nRMSE			sMAPE			BIAS_PCT		
		reg	clip	qp	reg	clip	qp	reg	clip	qp	reg	clip	qp
	arima	+1968 (+1023%)	+888 (+462%)	+296 (+154%)	+2338 (+1075%)	+1216 (+559%)	+489 (+225%)	+40 (+37%)	+112 (+11%)	+59 (+55%)	-35 (-29%)	+873 (+725%)	+229 (+190%)
	chronos	+1100 (+1509%)	+1089 (+862%)	+676 (+535%)	+2081 (+1305%)	+1257 (+796%)	+845 (+535%)	+27 (+23%)	+3 (+3%)	+39 (+33%)	+291 (+1391%)	+1108 (+5305%)	+663 (+3172%)
	deepar	+275 (+3391%)	+2016 (+1830%)	+940 (+854%)	+425 (+2987%)	+2576 (+1809%)	+1252 (+880%)	+43 (+40%)	+18 (+17%)	+53 (+49%)	+323 (+7755%)	+2042 (+49037%)	+923 (+22170%)
	holt_winters	+2045 (+1082%)	+988 (+523%)	+337 (+178%)	+2265 (+964%)	+1133 (+482%)	+420 (+179%)	+45 (+43%)	+17 (+16%)	+56 (+54%)	-81 (-77%)	+977 (+932%)	+291 (+277%)
	lightgbm	+1964 (+227%)	+729 (+84%)	+29 (+3%)	+2441 (+278%)	+1257 (+143%)	+500 (+57%)	+30 (+25%)	+5 (+5%)	+47 (+39%)	-58 (-67%)	+677 (+81%)	-73 (-9%)
	lightgbm_direct	+2207 (+247%)	+1106 (+124%)	+487 (+54%)	+2803 (+309%)	+1795 (+198%)	+955 (+105%)	+32 (+26%)	+5 (+4%)	+49 (+41%)	-54 (-6%)	+1047 (+121%)	+381 (+44%)
	nhits	+3825 (+3374%)	+1978 (+1745%)	+993 (+876%)	+4558 (+3183%)	+2915 (+2036%)	+1513 (+1111%)	+38 (+32%)	+10 (+8%)	+58 (+50%)	+150 (+2291%)	+197 (+30560%)	+953 (+14588%)
	prophet	+1943 (+1018%)	+1120 (+587%)	+743 (+389%)	+2078 (+911%)	+1224 (+536%)	+864 (+379%)	+45 (+43%)	+16 (+16%)	+60 (+58%)	+290 (+280%)	+1114 (+1073%)	+700 (+675%)
	random_forest	+2015 (+668%)	+883 (+293%)	+383 (+127%)	+2369 (+744%)	+1418 (+445%)	+787 (+247%)	+42 (+39%)	+12 (+12%)	+63 (+59%)	-208 (-83%)	+839 (+336%)	+287 (+115%)
	random_forest_direct	+2881 (+947%)	+1500 (+493%)	+845 (+278%)	+3350 (+1037%)	+1945 (+602%)	+1156 (+358%)	+45 (+42%)	+15 (+14%)	+66 (+63%)	+75 (+29%)	+1456 (+569%)	+749 (+292%)
	timesfm	+2386 (+2012%)	+1381 (+1165%)	+737 (+622%)	+2585 (+1749%)	+1556 (+1053%)	+900 (+609%)	+35 (+31%)	+12 (+11%)	+46 (+41%)	+394 (+2558%)	+1399 (+9084%)	+720 (+4678%)
WLS_struct		WMAPE			nRMSE			sMAPE			BIAS_PCT		
		reg	clip	qp	reg	clip	qp	reg	clip	qp	reg	clip	qp
	arima	+314 (+163%)	+95 (+49%)	+54 (+28%)	+384 (+177%)	+162 (+74%)	+121 (+56%)	+26 (+25%)	-0 (-0%)	+34 (+32%)	-89 (-74%)	+67 (+56%)	+10 (+9%)
	chronos	+480 (+380%)	+353 (+279%)	+331 (+262%)	+500 (+317%)	+367 (+232%)	+346 (+219%)	+5 (+4%)	-11 (-9%)	+8 (+7%)	+266 (+1271%)	+393 (+1881%)	+364 (+1744%)
	deepar	+723 (+666%)	+315 (+501%)	+107 (+462%)	+315 (+571%)	+116 (+433%)	+107 (+399%)	+10 (+16%)	+5 (+5%)	+10 (+19%)	+393 (+10023%)	+157 (+14373%)	+52 (+13126%)
	holt_winters	+410 (+217%)	+119 (+63%)	+48 (+26%)	+447 (+190%)	+144 (+61%)	+62 (+26%)	+33 (+32%)	+6 (+6%)	+42 (+40%)	-21 (-20%)	+102 (+98%)	+12 (+12%)
	lightgbm	+73 (+8%)	+126 (+15%)	+268 (+31%)	+179 (+20%)	+20 (-2%)	+173 (+18%)	+10 (+8%)	-1 (-1%)	+5 (+10%)	-256 (-43%)	+157 (+19%)	-310 (-37%)
	lightgbm_direct	-37 (-4%)	-144 (-16%)	-213 (-24%)	+65 (+7%)	-20 (-10%)	-87 (-10%)	+7 (+6%)	-2 (-3%)	+10 (+8%)	-284 (-33%)	-177 (-20%)	-253 (-29%)
	nhits	+2393 (+841%)	+1248 (+554%)	+594 (+487%)	+3182 (+784%)	+1942 (+567%)	+1242 (+505%)	+21 (+14%)	-2 (-2%)	+10 (+17%)	+324 (+5162%)	+159 (+10148%)	+397 (+8837%)
	prophet	+222 (+116%)	+85 (+45%)	+58 (+31%)	+234 (+102%)	+101 (+44%)	+75 (+33%)	+23 (+22%)	+2 (+2%)	+48 (+26%)	-50 (-48%)	+86 (+63%)	+48 (+46%)
	random_forest	+232 (+77%)	+34 (+11%)	-45 (-15%)	+303 (+95%)	+109 (+34%)	+72 (+7%)	+26 (+24%)	-2 (-2%)	+22 (+30%)	-208 (-83%)	-10 (-4%)	106 (+42%)
	random_forest_direct	+367 (+121%)	+177 (+58%)	+92 (+30%)	+453 (+140%)	+255 (+79%)	+167 (+52%)	+21 (+19%)	+4 (+4%)	+26 (+25%)	-44 (-17%)	+146 (+57%)	+45 (+18%)
	timesfm	+552 (+466%)	+395 (+333%)	+302 (+305%)	+565 (+382%)	+402 (+272%)	+370 (+250%)	+13 (+12%)	-2 (-2%)	+16 (+14%)	+270 (+1752%)	+427 (+2775%)	+387 (+2512%)

Слика 4.1: Структурно усклађивање (CSM): Промена перформанси након примене стабилних метода OLS и WLS_Struct кроз три различита приступа.

CSM | Reg vs Clip Only (Full Hierarchy, Base Level)
MinT_shrink & WLS_var (no QP available) | green = improvement, red = degradation

	WMAPE			nRMSE			sMAPE			BIAS_PCT		
	reg	clip	qp	reg	clip	qp	reg	clip	qp	reg	clip	qp
MinT_shrink	arima	+288 (+150%)	+145 (+76%)	+351 (+161%)	+231 (+106%)	+22 (+10%)	+7 (+3%)	-13 (-10%)	+130 (+108%)	-13 (-10%)	+6 (+30%)	+8 (+37%)
	chronos	+7 (+5%)	+5 (+4%)	+5 (+3%)	+5 (+3%)	+0 (+0%)	-8 (-7%)	+6 (+30%)	+8 (+37%)	+6 (+30%)	+4 (+19%)	+4 (+19%)
	deepar	+10 (+9%)	+8 (+7%)	+8 (+5%)	+6 (+4%)	+1 (+10%)	+1 (+1%)	+2 (+49%)	+4 (+99%)	+2 (+49%)	+4 (+99%)	+4 (+99%)
	holt_winters	-1 (-0%)	-8 (-4%)	-4 (-2%)	-9 (-4%)	-13 (-13%)	+4 (+4%)	-14 (-13%)	-6 (-6%)	-14 (-13%)	-6 (-6%)	-6 (-6%)
	lightgbm	+23 (+192%)	+98 (+97%)	+182 (+193%)	+182 (+99%)	+2 (+40%)	+2 (+22%)	-24 (-3%)	+98 (+96%)	-24 (-3%)	+98 (+96%)	+98 (+96%)
	lightgbm_direct	+417 (+47%)	+230 (+26%)	+417 (+46%)	+234 (+26%)	+25 (+21%)	+15 (+13%)	+24 (+3%)	+211 (+24%)	+24 (+3%)	+211 (+24%)	+211 (+24%)
	nhits	+19 (+22%)	+23 (+16%)	+23 (+16%)	+18 (+13%)	+1 (-1%)	+1 (+10%)	+1 (+10%)	+1 (+10%)	+1 (+10%)	+1 (+10%)	+1 (+10%)
	prophet	-1 (-0%)	-7 (-4%)	-4 (-2%)	-8 (-4%)	-11 (+10%)	+3 (+3%)	-11 (-11%)	-5 (-5%)	-11 (-11%)	-5 (-5%)	-5 (-5%)
	random_forest	+229 (+229%)	+90 (+90%)	+216 (+216%)	+87 (+87%)	+36 (+36%)	+15 (+15%)	-71 (-71%)	+98 (+98%)	-71 (-71%)	+98 (+98%)	+98 (+98%)
	random_forest_direct	-7 (-2%)	-10 (-3%)	-5 (-2%)	-8 (-3%)	+1 (+1%)	+0 (+0%)	-16 (-6%)	-13 (-5%)	+0 (+0%)	-16 (-6%)	-13 (-5%)
	timesfm	+7 (+6%)	+5 (+4%)	+5 (+4%)	+4 (+3%)	+4 (+4%)	+4 (+4%)	+6 (+36%)	+8 (+51%)	+4 (+36%)	+6 (+51%)	+8 (+51%)
WLS_var	arima	+18 (+9%)	+5 (+3%)	+21 (+10%)	+8 (+4%)	+1 (+1%)	-0 (-0%)	-9 (-7%)	+4 (+3%)	-9 (-7%)	+4 (+3%)	+4 (+3%)
	chronos	+7 (+5%)	+5 (+5%)	+5 (+4%)	+5 (+4%)	+0 (+0%)	+0 (+0%)	+6 (+46%)	+10 (+47%)	+0 (+0%)	+6 (+46%)	+10 (+47%)
	deepar	+7 (+6%)	+6 (+6%)	+5 (+4%)	+5 (+4%)	+11 (+10%)	+11 (+10%)	+2 (+48%)	+2 (+48%)	+11 (+10%)	+2 (+48%)	+2 (+48%)
	holt_winters	-0 (-0%)	-2 (-2%)	-1 (-1%)	-3 (-1%)	-15 (+14%)	+3 (+3%)	-7 (-7%)	-4 (-4%)	-15 (+14%)	+3 (+3%)	-4 (-4%)
	lightgbm	-120 (-140%)	-124 (-140%)	-119 (-140%)	-123 (-140%)	-1 (-1%)	-2 (-1%)	-133 (-16%)	-128 (-15%)	-1 (-1%)	-133 (-16%)	-128 (-15%)
	lightgbm_direct	-109 (-12%)	-114 (-13%)	-108 (-12%)	-113 (-12%)	-1 (-1%)	-1 (-1%)	-123 (-14%)	-118 (-14%)	-1 (-1%)	-123 (-14%)	-118 (-14%)
	nhits	+9 (+8%)	+9 (+8%)	+9 (+8%)	+9 (+8%)	-1 (-1%)	-1 (-1%)	+13 (+192%)	+13 (+194%)	-1 (-1%)	+13 (+192%)	+13 (+194%)
	prophet	+1 (+0%)	-1 (-1%)	+0 (+0%)	-1 (-1%)	+13 (+13%)	+3 (+3%)	-3 (-3%)	-1 (-1%)	+13 (+13%)	+3 (+3%)	-1 (-1%)
	random_forest	-8 (-3%)	-9 (-3%)	-8 (-2%)	-9 (-3%)	+0 (+0%)	+0 (+0%)	-12 (-5%)	-11 (-4%)	+0 (+0%)	-12 (-5%)	-11 (-4%)
	random_forest_direct	-6 (-2%)	-7 (-2%)	-6 (-2%)	-7 (-2%)	+0 (+0%)	+0 (+0%)	-10 (-4%)	-9 (-3%)	+0 (+0%)	-10 (-4%)	-9 (-3%)
	timesfm	+1 (+6%)	+1 (+6%)	+1 (+4%)	+1 (+4%)	+7 (+4%)	+6 (+3%)	+11 (+70%)	+11 (+70%)	+7 (+4%)	+11 (+70%)	+11 (+70%)

Слика 4.2: Структурно усклађивање (CSM): Промена перформанси након примене емпиријских метода MinT_shrink и WLS_var (QP оптимизација изостављена).

Структурно усклађивање

Анализа топлотних мапа за структурну димензију (слике 4.1 и 4.2) открива један од најважнијих увида овог рада: **структурно усклађивање по правилу драстично деградира тачност прогноза на базном нивоу**.

Детерминистичка метода *OLS* узрокује велику деградацију. На пример, код *DeepAR* модела, *OLS* у неограниченом (*reg*) режиму повећава *WMAPE* за екстремних +3391%, док код *Chronos* модела грешка расте за +1509%. Овакав резултат је математички очекиван у системима са 20.000 серија: матрица *OLS* претпоставља да све серије имају исту дисперзију грешке, те слепо дистрибуира корекције са виших нивоа на ниже, потпуно уништавајући квалитетне локалне сигнале.

Примена методе *WLS_Struct* значајно ублажава ову деградацију (нпр. *WMAPE* за *DeepAR* расте за +666%), али и даље не доноси побољшање. Слично се понаша и *MinT_Shrink* метода, код које квалитетни базни модели попут *Chronos*-а и *DeepAR*-а бележе благи пад тачности (око +5% до +9%).

Једини изузетак који показује висок ниво стабилности јесте метода *WLS_Var*. Код ове методе, одлични модели задржавају своје перформансе уз минималну деградацију (нпр. *Chronos* *WMAPE* + 5%), док изразито лоши модели попут *LightGBM*-а доживљавају општро побољшање (−14%). Овај налаз сугерише да је за успешно структурно усклађивање испрекиданих временских серија есенцијално скалирање корекција искључиво на основу историјске дисперзије сваке појединачне серије, чиме се спречава преливање туђег шума на квалитетне прогнозе.

Утицај дубине хијерархије:

Додатни експеримент са редукованом структурном хијерархијом (Слика 4.3) показао је да уклањање нивоа *кућаци*, *производна хијерархија 2* и *производна хијерархија 3* ублажава негативне ефекте усклађивања. На пример, деградација *ARIMA* модела при *OLS*-у пада са +1023% на +341%. Мањи број нивоа значи мањи број степени слободе и мање веза у матрици сабирања **S**, чиме се ограничава могућност алгорита да прилагођава базне серије како би задовољио збирове на удаљеним нивоима.

ГЛАВА 4. РЕЗУЛТАТИ И ДИСКУСИЈА

CSM Full vs Reduced Hierarchy (approach=reg, Base Level)									
Columns: full / subset green = improvement vs base model									
Min_T_shrink	WMAPE		nRMSE		sMAPE		BIAS_PCT		
	arima	+288 (+150%)	+27 (+14%)	+351 (+161%)	+40 (+18%)	+22 (+21%)	+6 (+6%)	-13 (-10%)	-10 (-8%)
	chronos	+7 (+5%)	+3 (+2%)	+5 (+3%)	+3 (+2%)	+0 (+0%)	+1 (+1%)	+6 (+30%)	+4 (+17%)
	deepar	+10 (+9%)	+5 (+5%)	+8 (+5%)	+3 (+2%)	+11 (+10%)	+12 (+11%)	+2 (+49%)	-4 (-92%)
	holt_winters	-1 (-0%)	-0 (-0%)	-4 (-2%)	-2 (-1%)	+13 (+13%)	+14 (+13%)	-14 (-13%)	-4 (-4%)
	lightgbm	+1662 (+192%)	+1609 (+186%)	+1699 (+193%)	+1646 (+187%)	+48 (+40%)	+48 (+40%)	-24 (-3%)	-49 (-6%)
	lightgbm_direct	+417 (+47%)	+107 (+12%)	+417 (+46%)	+109 (+12%)	+25 (+21%)	+11 (+9%)	+24 (+3%)	+60 (+7%)
	nhits	+24 (+22%)	+19 (+16%)	+23 (+16%)	+17 (+12%)	-1 (-1%)	-1 (-1%)	+22 (+33%)	+16 (+24%)
	prophet	-1 (-0%)	-4 (-0%)	-4 (-2%)	-1 (-1%)	+11 (+10%)	+12 (+12%)	-11 (-11%)	-3 (-3%)
	random_forest	+691 (+229%)	-3 (-1%)	+689 (+216%)	-3 (-1%)	+38 (+36%)	+0 (+0%)	-176 (-71%)	-6 (-2%)
	random_forest_direct	-7 (-2%)	-1 (-0%)	-5 (-2%)	-1 (-0%)	+1 (+1%)	+0 (+0%)	-16 (-6%)	-2 (-1%)
	timesfm	+7 (+6%)	+3 (+2%)	+6 (+4%)	+2 (+2%)	+4 (+4%)	+5 (+4%)	+6 (+36%)	+3 (+18%)
	full	subset	full	subset	full	subset	full	subset	
OLS	WMAPE		nRMSE		sMAPE		BIAS_PCT		
	arima	+1968 (+1023%)	+657 (+341%)	+2338 (+1075%)	+776 (+357%)	+40 (+37%)	+34 (+32%)	-35 (-29%)	+182 (+151%)
	chronos	+1906 (+1509%)	+1017 (+805%)	+2061 (+1305%)	+1072 (+679%)	+27 (+23%)	+16 (+13%)	+291 (+1391%)	+265 (+1266%)
	deepar	+3135 (+3391%)	+1922 (+1745%)	+4252 (+2987%)	+2224 (+1562%)	+43 (+40%)	+22 (+22%)	+23 (+775%)	+368 (+9349%)
	holt_winters	+2045 (+1082%)	+664 (+352%)	+2265 (+964%)	+717 (+305%)	+45 (+43%)	+39 (+37%)	-81 (-77%)	+225 (+215%)
	lightgbm	+1964 (+227%)	+549 (+63%)	+2441 (+278%)	+735 (+84%)	+30 (+25%)	+13 (+11%)	-558 (-67%)	-516 (-62%)
	lightgbm_direct	+2207 (+247%)	+725 (+81%)	+2803 (+309%)	+1040 (+115%)	+32 (+26%)	+18 (+15%)	-54 (-6%)	-54 (-6%)
	nhits	+825 (+3374%)	+1016 (+1425%)	+1458 (+3183%)	+1602 (+314%)	+36 (+32%)	+24 (+20%)	+150 (+2291%)	+368 (+8953%)
	prophet	+1943 (+1018%)	+1011 (+530%)	+2078 (+911%)	+1048 (+459%)	+45 (+43%)	+33 (+32%)	+290 (+280%)	+453 (+436%)
	random_forest	+2015 (+668%)	+938 (+311%)	+2369 (+744%)	+735 (+346%)	+42 (+39%)	+33 (+31%)	-208 (-83%)	-94 (-37%)
	random_forest_direct	+2881 (+947%)	+1342 (+441%)	+3350 (+1037%)	+1562 (+484%)	+45 (+42%)	+32 (+30%)	+75 (+29%)	+327 (+128%)
	timesfm	+2316 (+2012%)	+968 (+816%)	+2405 (+1749%)	+1063 (+720%)	+35 (+31%)	+19 (+17%)	+394 (+2558%)	+339 (+2201%)
	full	subset	full	subset	full	subset	full	subset	
WLS_struct	WMAPE		nRMSE		sMAPE		BIAS_PCT		
	arima	+314 (+163%)	+183 (+95%)	+384 (+177%)	+219 (+100%)	+26 (+25%)	+27 (+25%)	-89 (-74%)	-9 (-7%)
	chronos	+480 (+380%)	+501 (+297%)	+500 (+317%)	+501 (+213%)	+5 (+4%)	+1 (+1%)	+266 (+1271%)	+516 (+2470%)
	deepar	+733 (+666%)	+751 (+682%)	+813 (+571%)	+797 (+560%)	+17 (+16%)	+12 (+13%)	+417 (+10023%)	+684 (+16418%)
	holt_winters	+410 (+217%)	+229 (+121%)	+447 (+190%)	+243 (+104%)	+33 (+32%)	+36 (+35%)	-21 (-20%)	+44 (+42%)
	lightgbm	+73 (+8%)	-125 (-14%)	+179 (+20%)	-47 (-5%)	+10 (+8%)	+7 (+6%)	-356 (-43%)	-335 (-40%)
	lightgbm_direct	-37 (-4%)	-119 (-13%)	+65 (+7%)	-46 (-5%)	+7 (+6%)	+3 (+2%)	-284 (-33%)	-230 (-27%)
	nhits	+953 (+841%)	+674 (+594%)	+1123 (+784%)	+740 (+517%)	+16 (+14%)	+5 (+4%)	+337 (+5162%)	+670 (+10259%)
	prophet	+222 (+116%)	+119 (+62%)	+234 (+102%)	+131 (+58%)	+23 (+22%)	+18 (+17%)	-50 (-48%)	-0 (-0%)
	random_forest	+232 (+77%)	+128 (+43%)	+303 (+95%)	+194 (+61%)	+26 (+24%)	+31 (+29%)	-208 (-83%)	-116 (-46%)
	random_forest_direct	+367 (+121%)	+184 (+60%)	+453 (+140%)	+233 (+72%)	+21 (+19%)	+16 (+15%)	-44 (-17%)	-42 (-16%)
	timesfm	+552 (+466%)	+401 (+389%)	+555 (+382%)	+401 (+312%)	+13 (+12%)	+6 (+5%)	+210 (+1752%)	+487 (+3160%)
	full	subset	full	subset	full	subset	full	subset	
WLS_var	WMAPE		nRMSE		sMAPE		BIAS_PCT		
	arima	+18 (+9%)	+1 (+0%)	+21 (+10%)	+1 (+0%)	+1 (+1%)	+1 (+0%)	-9 (-7%)	-2 (-2%)
	chronos	+3 (+5%)	+3 (+2%)	+5 (+4%)	+2 (+2%)	+0 (+0%)	+1 (+1%)	+10 (+46%)	+4 (+18%)
	deepar	+7 (+6%)	+3 (+2%)	+5 (+4%)	+2 (+1%)	+11 (+10%)	+12 (+11%)	+2 (+48%)	-4 (-94%)
	holt_winters	-0 (-0%)	-0 (-0%)	-1 (-1%)	-1 (-0%)	+15 (+14%)	+15 (+14%)	-7 (-7%)	-2 (-2%)
	lightgbm	-120 (-14%)	-20 (-2%)	-119 (-14%)	-19 (-2%)	-1 (-1%)	-0 (-0%)	-133 (-16%)	-21 (-3%)
	lightgbm_direct	-109 (-12%)	-24 (-3%)	-108 (-12%)	-24 (-3%)	-1 (-1%)	+0 (+0%)	-123 (-14%)	-27 (-3%)
	nhits	+9 (+8%)	+5 (+4%)	+8 (+6%)	+5 (+3%)	-1 (-1%)	+0 (+0%)	+13 (+192%)	+7 (+100%)
	prophet	+1 (+0%)	+0 (+0%)	+0 (+0%)	-0 (-0%)	+13 (+13%)	+13 (+13%)	-3 (-3%)	-1 (-1%)
	random_forest	-8 (-3%)	-3 (-1%)	-8 (-2%)	-2 (-1%)	+0 (+0%)	+0 (+0%)	-12 (-5%)	-4 (-2%)
	random_forest_direct	-6 (-2%)	-1 (-0%)	-6 (-2%)	-1 (-0%)	+0 (+0%)	+0 (+0%)	-10 (-4%)	-1 (-1%)
	timesfm	+8 (+6%)	+3 (+2%)	+7 (+4%)	+2 (+2%)	+4 (+4%)	+5 (+4%)	+11 (+70%)	+4 (+24%)
	full	subset	full	subset	full	subset	full	subset	

Слика 4.3: Утицај дубине структурне хијерархије (Пуна наспрам Редуковане) на деградацију базних прогноза.

Временско усклађивање

За разлику од структурног усклађивања, темпорална димензија (прелазак са месечне на кварталну агрегацију) манифестује потпуно другачије обрасце понашања.

Прво, резултати за сва три приступа (*reg*, *clip*, *qp*) приказани на слици 4.4 су идентични код темпоралног усклађивања. Ово потврђује да агрегација из месечне у кварталну фреквенцију ретко или готово никада не генерише математички негативне прогнозе, те интервенције одсецања негативних вредности и квадратног програмирања нису имале никакав ефекат на систем.

Друго, темпорално усклађивање генерише снажна побољшања искључиво код најлошијих модела (нпр. *LightGBM* бележи пад WMAPE метрике за -36% до -60%). Овде је важно разумети саму механику алгорита: MinT користи независно генерисане кварталне прогнозе како би кориговао месечне. Код лоших модела са високим нивоом шума, ове независне кварталне прогнозе су често релативно стабилније од месечних. Када алгоритам дистрибуира ту кварталну корекцију назад на месечни ниво, он ефективно врши пост-хок регуларизацију и смањује укупну грешку. Код детерминистичких метода (*OLS*) ова дистрибуција је равномерна, док емпиријска метода *WLS_Var* ово ради интелигентно, тј. асиметрично коригује само оне месеце за које је модел историјски имао високу несигурност. Управо зато метода *WLS_Var* доноси и највеће побољшање за лоше моделе.

С друге стране, најбољи предиктивни модели (*DeepAR*, *N-HITS*, *TimesFM*) доживљавају деградацију (повећање WMAPE за $+10\%$ до $+20\%$) након темпоралног усклађивања. Ови модели већ генеришу оштре и прецизне локалне сигнале на месечном нивоу. Када алгоритам покуша да их усклади са њиховим кварталним прогнозама, он њихов квалитетан локални сигнал вештачки дистрибуира унутар квартала. Коначни закључак је да ефикасност темпоралног усклађивања директно зависи од количине шума у базним прогнозама: оно успешно регуларизује лоше калибрисане моделе, али истовремено доноси деградацију алгоритмима чији је базни сигнал већ оптималан. Стога се хијерархијско усклађивање у врхунским предиктивним системима не сме нужно посматрати као алат за смањење локалне грешке, већ примарно као административни механизам који обезбеђује пословну кохерентност целог система (која је неопходна за корпоративно извештавање и планирање буџета), при чему та кохерентност долази по цену благог пада прецизности на најнижем

нивоу.

4.3 Валидација и нарушавање хијерархијске кохерентности

Као што је теоријски дефинисано у глави 2, основни циљ хијерархијског усклађивања је постизање агрегационе кохерентности система (збир прогноза деце мора бити једнак прогнози родитеља). Интерна евалуација на тест скупу потврдила је да неограничено усклађивање, као и усклађивање уз формална QP ограничења, остварују савршену структурну и темпоралну кохерентност (100% успешних провера над свим чворовима).

Међутим, примена хеуристичког одсецања негативних вредности подразумева нелинеарну трансформацију која неминовно нарушава ову математичку структуру. Како би се квантификовала цена овог инжењерског компромиса, спроведена је анализа. За сваки родитељски чвор упоређена је његова усклађена, а затим одсечена вредност са збиром усклађених, а затим одсечених вредности његових потомака са базног нивоа. Резултати ове анализе приказани су у табели 4.2. За сваки ниво агрегације посматрана су два кључна показатеља. Први показатељ, *ушницај*, представља проценат чворова на датом нивоу код којих је дошло до неслагања између родитеља и деце услед одсецања. Други показатељ *нарушавање кохерентности* мери просечну процентуалну разлику између вредности агрегираног извештаја на том нивоу и физичког збира његових базних серија. Кључни налази ове анализе, указују на следеће:

- **Укупно нарушавање система:** Агрегатна процентуална грешка кохерентности на нивоу целог система износи **6.00%**. Ово значи да се, у просеку, извештаји на вишим агрегационим нивоима разликују за 6% од физичког збира базних резервних делова услед одсецања негативних прогноза.
- **Дистрибуција по нивоима:** Губитак кохерентности је израженији на највишим нивоима агрегације (укупни *укупно* ниво 6.92%, *пословна област* 6.80%). Спуштањем кроз производну хијерархију ка грануларнијим нивоима, ово одступање природно опада (*производна хијерархија*

3 ниво бележи одступање од 4.82%). Овај феномен је логичан, јер виши нивои сабирају већи број базних серија, чиме се кумулативни ефекат одсечених нула додатно појачава.

Табела 4.2: Просечно процентуално нарушавање кохерентности услед хеуристичког одсецања негативних вредности, приказано по структурним нивоима.

Ниво	Утицај (% чворова)	Нарушавање кохерентности (%)
<i>укупно</i>	66.7%	6.92%
<i>пословна област</i>	47.6%	6.80%
<i>локација</i>	48.6%	6.55%
<i>производна хијерархија 1</i>	36.1%	6.04%
<i>кућац</i>	38.1%	5.81%
<i>производна хијерархија 2</i>	27.3%	5.11%
<i>производна хијерархија 3</i>	23.7%	4.82%
агрегатно	-	6.00%

Поред агрегационих нивоа, ниво нарушавања кохерентности представља директну последицу интеракције између квалитета базног модела и коришћене матрице усклађивања. Ова динамика приказана је у табели 4.3.

Табела 4.3: Процентуално нарушавање кохерентности услед хеуристичког одсецања негативних вредности, приказано по базним моделима и примењеним методама усклађивања.

Модел	OLS	WLS_Struct	MinT_Shrink	WLS_Var
<i>DeepAR</i>	15.24	1.11	0.05	0.001
<i>N-HITS</i>	17.72	2.07	0.28	0.01
<i>TimesFM</i>	7.78	0.74	0.06	0.004
<i>Chronos</i>	6.04	0.57	0.03	0.003
<i>Holt-Winters</i>	9.15	2.53	0.67	0.48
<i>Prophet</i>	8.23	1.29	0.48	0.29
<i>ARIMA</i>	12.73	2.13	15.71	0.50
<i>Random Forest</i>	12.58	2.53	48.32	0.06
<i>Random Forest (dir.)</i>	16.59	2.52	0.11	0.04
<i>LightGBM</i>	13.37	2.19	81.91	0.16
<i>LightGBM (dir.)</i>	13.91	1.77	54.38	0.26
просек	11.21	1.77	18.36	0.16

- **Детерминистички приступ (OLS):** Ова метода је изразито нестабилна на свим моделима. Она слепо дистрибуира корекције и генерише велике негативне вредности чијим се одсецањем кохерентност трајно нарушава (конзистентно између 6% и 17%).
- **Структурно скалирање (WLS_Struct):** Ова метода представља стабилну средину. Коришћењем саме хијерархијске структуре за апроксимацију несигурности, она значајно ублажава екстреме у односу на *OLS*. Без обзира на квалитет базног модела, задржава губитак кохерентности у уским и пословно прихватљивим границама (између 0.5% и 2.5%).
- **Емпиријски приступи (WLS_Var и MinT_Shrink):** Метода *WLS_Var* доказује своју супериорност јер интелигентно скалира корекције према стварној дисперзији серија, чиме се потреба за одсецањем негативних вредности и губитак кохерентности своде на ниво занемарљиве грешке (често испод 0.01%). Са друге стране, метода *MinT_Shrink* показује екстремну осетљивост на одударајуће податке: код квалитетних модела не квари збирове, али у комбинацији са изразито пристрасним *LightGBM*-ом доводи до драматичног пада кохерентности (преко 80% грешке).

4.4 Дискусија добијених резултата

Темељна евалуација на подацима из домена резервних делова преиспитује традиционалне претпоставке о хијерархијском усклађивању. Експеримент недвосмислено показује да усклађивање није универзално решење за побољшање предиктивне тачности. На подацима са великим уделом испрекиданих серија, агрегација уноси значајан ниво шума. Уколико се тај шум слепо дистрибуира назад на базни ниво (као код методе *OLS*), долази до велике деградације локалних сигнала. Када систем користи врхунске алгоритме дубоког учења или вишенаменске претрениране моделе, њихове базне прогнозе су већ толико добро калибрисане да им линеарно усклађивање са вишим нивоима готово уопште не помаже, већ их благо погоршава.

Значајан аспект дискусије односи се на утицај дубине, односно дизајна саме структурне хијерархије и на квалитет усклађених прогноза. Поређење

пуне и редуковане хијерархије недвосмислено показује да комплекснија структура, са већим бројем међунивоа (агрегационих погледа), уноси додатну грешку у систем. Сваки нови ниво уводи додатна ограничења која базне серије морају да задовоље, што алгоритам приморава на непотребну адаптацију локалних сигнала како би се поклопили са удаљеним збировима. Смањењем броја агрегационих погледа (редукована хијерархија), ефективно се смањује број степени слободе у пројекционој матрици, чиме се значајно ублажава деградација базних прогноза. Ово потврђује да дизајн хијерархије не сме бити вођен искључиво организационим корпоративним шемама, већ се мора пажљиво балансирати како алгебарска ограничења не би преоптеретила предиктивни систем.

Ефекат темпоралног усклађивања захтева посебно опрезну интерпретацију. Уместо простог математичког усредњавања, алгоритам темпоралног усклађивања користи независно генерисане агрегиране (кварталне) прогнозе како би кориговао базне (месечне) сигнале. Детаљна анализа открива да овај процес суштински функционише као механизам пост-хок регуларизације. Дистрибуцијом кварталних вредности наниже, алгоритам ефективно ублажава екстремне локалне осцилације и велику пристрасност лоших модела, чиме им стварно и значајно смањује укупну грешку. Међутим, инхерентна цена ове регуларизације је вештачко погоршавање већ прецизних локалних сигнала код висококвалитетних модела. Због тога, закључак да темпорално усклађивање побољшава систем у просеку заправо представља последицу екстремних поправки најлошијих модела. У реалности, напредним алгоритмима ова врста темпоралне интервенције не доноси корист, већ доводи до благог пада њихове иницијалне прецизности.

Коначно, анализа доменских ограничења потврђује да је у индустријској пракси инжењерски компромис често бољи од строге математике. Иако формално квадратно програмирање (QP) гарантује кохерентност без одсецања негативних вредности, његова рачунарска нестабилност чини га неупотребљивим за системе са 20.000 серија. Хеуристичко одсецање негативних вредности на нулу успешно решава овај проблем уз занемарљиву цену: упарен са интелигентном методом WLS_Var , он нарушава математичку кохерентност система за мање од 0.5%, што је у пословном извештавању углавном потпуно прихватљива маргина грешке.

4.5 Идентификација најбољих пракси за дати проблем

На основу синтезе резултата, дефинисане су конкретне инжењерске препоруке за пројектовање предиктивних система у описаном домену:

1. **Избор алгоритама:** Приоритет треба дати моделима дубоког учења (*DeepAR*) или вишенаменским претренираним моделима (*Chronos*, *TimesFM*) посебно ако нам је потребно да брзо дођемо до иницијалних резултата. Класична стабла одлучивања (*LightGBM*, *Random Forest*) не треба користити на скупу са великим уделом испрекиданих временских серија без претходне промене функције циља и конструкције карактеристика.
2. **Стратегија примене усклађивања:** Одлука о димензији усклађивања мора бити директно вођена пословним приоритетима:
 - Уколико је **максимална тачност на базном нивоу** апсолутни приоритет (нпр. за аутоматизовано наручивање делова), а користе се квалитетни модели, хијерархијско усклађивање треба **потпуно избегавати**.
 - Уколико су **кохерентни извештаји** неопходни за менаџмент, применити структурно усклађивање, али искључиво кроз методу ***WLS_Var***, и уз пажљив избор хијерархија, како би се сачувао интегритет добрих базних модела.
 - Темпорално усклађивање треба стратешки користити искључиво као механизам пост-хок регуларизације у системима где се нужно користе лошији, нестабилни модели са високом дисперзијом. За добре моделе, темпорално усклађивање не доноси побољшања.
3. **Третман негативних прогноза:** Уместо нумерички нестабилних QR решавача, препоручује се неограничено усклађивање праћено хеуристичким одсецањем негативних вредности на нулу. Упарен са методом ***WLS_Var***, овај приступ нуди максималну стабилност уз занемарљив губитак кохерентности.

4. **Дизајн хијерархије:** Комплексност агрегације треба свести на апсолутни минимум. Увођење непотребних међунивоа у структурну хијерархију повећава број степени слободе и доводи до прекомерних, штетних корекција базних серија.

Глава 5

Закључак

5.1 Синтеза кључних налаза истраживања

У овом мастер раду спроведена је свеобухватна анализа утицаја хијерархијског усклађивања на прецизност предвиђања временских серија, мотивисана индустријским проблемом предвиђања потражње резервних делова. Специфичност овог домена огледа се у високој димензионалности, сложеној структури агрегације и великом уделу испрекиданих временских серија.

Најзначајнији допринос овог рада лежи у побијању претпоставке да хијерархијско усклађивање аутоматски побољшава локалну прецизност модела. Резултати недвосмислено доказују да врхунски алгоритми, посебно вишестепенски претренирани модели (*Chronos*, *TimesFM*) и вероватносне неуронске мреже (*DeepAR*), екстрахују максимум релевантних информација директно из сирових података. Накнадно линеарно усклађивање таквих добро калибрисаних локалних сигнала са независним агрегираним нивоима кроз алгоритме MinT не доноси корист, већ често уноси грешку. За високо прецизне моделе, хијерархијско усклађивање се стога мора посматрати превасходно као неопходан административни механизам за корпоративно извештавање, а не као алатка за минимизацију локалне грешке.

Значајан налаз ове студије је и директан утицај самог дизајна хијерархије на резултате усклађивања. Показано је да комплекснија структура са већим бројем агрегационих погледа уноси додатну грешку у систем. Редукована хијерархија (са мањим бројем нивоа) показала се као боља, јер смањење броја степени слободе у матрици сабирања ефективно ограничава нежељено прилагођавање базних серија.

Поред тога, рад се дотакао и решавања конфликта између строге математичке кохерентности и пословних ограничења (ненегативности). Показали смо да је у високодимензионалним системима замена нестабилне формалне оптимизације (QP) рачунарски јефтиним хеуристикама (одсецање негативних вредности на нулу) неопходна за физичку употребљивост прогноза. Упарен са методама које поштују историјску дисперзију (попут *WLS_Var*), овај инжењерски компромис чува агрегациони интегритет система уз занемарљив губитак математичке кохерентности.

5.2 Ограничења спроведене анализе

Као и свако истраживање, и ово има своја ограничења која треба узети у обзир приликом интерпретације резултата.

Прво ограничење односи се на рачунарске и математичке ресурсе при раду са високодимензионалним системима. Због екстремне меморијске захтевности, није било могуће извршити експлицитну алокацију велике матрице за истовремено (заједничко) структурно и темпорално усклађивање над 20.000 серија. Такође, због природе података где је број серија знатно већи од броја доступних временских корака ($n \gg T$), математички је немогуће инвертовати пуну емпиријску коваријациону матрицу. Због тога није било могуће тестирати класични алгоритам MinT.

Друго ограничење везано је за моделе класичног машинског учења (*LightGBM*, *Random Forest*). Ови алгоритми су намерно тренирани искључиво са историјским и календарским карактеристикама, уз стандардну MSE функцију циља. Разлог за овако скроман избор улазних података било је фер поређење са статистичким и вишенаменским претренираним моделима, којима је доступна искључиво историја циљне променљиве. Иако је тиме спречено да се моделима заснованим на стаблима даје предност кроз екстерне информације, то уједно значи да приказане перформансе ових алгоритама у раду не представљају њихов теоријски максимум.

Коначно, важно је нагласити да су сви закључци изведени на специфичном скупу података из домена потражње резервних делова, који одликује екстремна испрекиданост и велики број нултих вредности. Ови резултати се не могу аутоматски генерализовати; постоји велика вероватноћа да би на „глатким” временским серијама са стабилним и великим волуменом продаје (нпр.

у класичној малопродаји робе широке потрошње) линеарно усклађивање кроз алгоритме MinT дало значајно другачије, вероватно повољније резултате.

5.3 Препоруке за будућа истраживања и практичну примену

Претходно наведена ограничења директно инспиришу и дефинишу правце за будућа истраживања. Када је реч о предиктивном моделовању, логичан наредни корак био би развој модела заснованих на стаблима кроз имплементацију прилагођених функција губитка и увођење богатијег скупа екстерних карактеристика. Ово би омогућило да се испита да ли ти алгоритми могу да парирају вишенаменским претренираним моделима када нису ограничени. Такође, репликација ове експерименталне поставке на скуповима „глатких” временских серија високог волумена омогућила би компаративну анализу понашања алгоритама MinT у суштински различитим тржишним условима.

У домену алгоритмике усклађивања, будућа истраживања би се могла фокусирати на технике смањења димензионалности које би омогућиле апроксимацију класичног MinT приступа чак и када је $N \gg T$, као и на алгоритамску оптимизацију меморијских захтева за заједничко структурно-временско усклађивање. Код вероватносних модела фокус би се могао померити на вероватносно усклађивање, где би се кроз хијерархију уместо тачкастих прогноза пројектовале целе вероватносне расподеле. Такође, постоји огроман потенцијал у интеграцији најновијих парадигми неуронског усклађивања (енг. *neural reconciliation*) [15]. Уместо ослањања на фиксне линеарне пројекције (попут матрица MinT), ови савремени приступи користе архитектуре попут аутоенкодера или потпуно повезаних мрежа како би директно из података научили оптималне, нелинеарне пресликавајуће функције. Имплементација оваквих нелинеарних алгоритама представља један од најперспективнијих праваца тренутно.

Резултати истраживања дају увид у то како избор методе усклађивања утиче на тачност предвиђања, математичку кохерентност и рачунарску стабилност, што може бити корисно за пројектовање робусних предиктивних система.

Библиографија

- [1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.
- [2] George Athanasopoulos, Roman A Ahmed, and Rob J Hyndman. Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25(1):146–166, 2009.
- [3] George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74, 2017.
- [4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [5] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6989–6997, 2023.
- [6] Byron J Dangerfield and John S Morris. Bottom-up or top-down: Summary versus item forecasting. *International Journal of Forecasting*, 8(2):233–241, 1992.
- [7] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.
- [8] Charles W Gross and Jeffrey E Sohl. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 6(4):233–254, 1990.

- [9] Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 2011.
- [10] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 3rd edition, 2021.
- [11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Liu Tie-Yan. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, volume 30, 2017.
- [12] Nikolaos Kourentzes and George Athanasopoulos. Cross-temporal coherent forecasts for australian tourism. *Annals of Tourism Research*, 75:393–409, 2019.
- [13] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.
- [14] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.
- [15] Paul Mancuso, Vincenzo Picillo, and Michele Rubattu. Machine learning for hierarchical forecasting. *International Journal of Forecasting*, 37(4):1670–1686, 2021.
- [16] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.
- [17] Kin G Olivares, Cristian Challu, Federico Garza, Max Mergenthaler Canseco, and Artur Dubrawski. Hierarchicalforecast: A reference framework for hierarchical forecasting in python. *arXiv preprint arXiv:2207.03517*, 2022.
- [18] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2020.

- [19] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.
- [20] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.
- [21] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
- [22] Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.