АВТОМАТИЗИРАНА АНАЛИТИКА: ПРОГНОЗИРАНЕ НА ВРЕМЕВИ РЕДОВЕ В ИКОНОМИКАТА И БИЗНЕСА

гл. ас. д-р Антон Антонов Герунов
Софийски университет „Св. Климент Охридски“

AUTOMATING ANALYTICS: FORECASTING TIME SERIES IN ECONOMICS AND BUSINESS

Chief Assist. Prof. Anton Antonov Gerunov, PhD
Sofia University „St. Kliment Ohridski”

Abstract: *This paper demonstrates a feasible approach to fit an automated forecasting algorithm on four crucial economic time series from the Bulgarian economy. We use data on GDP growth, inflation, unemployment, and interest rates and estimate a large number of possible models. The best ones are selected by taking recourse to the Akaike Information Criterion. The optimal ARIMA models are studied and commented. Forecast accuracy metrics are presented and a few major conclusions and possible model applications are outlined. The paper concludes with directions for further research.*

Key words: *Automated analytics, forecasting, time series, ARIMA, business forecasting*

I. Въведение

Анализът на големи масиви от данни е бързо растящо поле в сферата на бизнес и икономическите анализи в последните години. С нарастващите възможности на организациите от публичния и частния сектор да набират все по-голямо количество различни данни в реално време се увеличават и предизвикателствата при тяхното използване за нуждите на бизнеса или публичната политика. Подобни масиви представят специфични изисквания за съхранението, обработката и анализа на данните в тях, които отиват отвъд стандартните методи, използвани днес. Големи по размер организации произвеждат стотици или дори хиляди индикатори за тяхната оперативна дейност, логистика, финанси, продажби и управление на ресурсите. Ключово при наличието на толкова много информация е нейното успешно използване с цел отключване на икономическа стойност. От друга страна организациите могат да се възползват в пълна степен от тези възможности само и единствено ако осигурят адекватно ресурсно обезпечаване на аналитичните си дейности. Предвид недостига на обучена работна сила и необходимостта от значителни инвестиции за нейното наемане и развитие, разходите за аналитични дейности биха могли да се окажат забранително високи.

От друга страна много организации все пак имат нужда да анализират наличните си данни за целите на планирането и логистичното обезпечаване. По същия начин много административни звена и регулатори имат нужда от аналитична информация в помощ на провеждането на ефективна публична политика. Всичко това подчертава необходимостта от създаване на широкомащабни аналитични продукти при условия

на значителни ресурсни ограничения. Настоящата разработка изследва едно възможно решение – автоматизиране на аналитичния процес с особен фокус върху прогнозирането на променливи, които се изменят във времето (времеви редове). Подобен подход позволява скалируемо прогнозиране на хиляди променливи с относително висока точност на оценките в рамките на малък период от време и при ниски разходи. Първоначално правим кратък преглед на литературата във връзка с анализа на времеви редове, а след това демонстрираме автоматичен алгоритъм за прогнозиране на четири ключови икономически променливи – БВП, инфлация, безработица и лихвени проценти. Разработката завършва с няколко основни изводи и потенциални идеи за бъдещи изследвания.

II. Преглед на литературата и актуалност на темата

Поради своята важност за нуждите на планирането, прогнозирането се радва на значително внимание в статистическата, икономическата и бизнес литературата. Прилагат се множество подходи за прогнозиране, които варират от сравнително базисни екстраполативни методи, през относително по-сложни подходи на авторегресия и плъзгащи се средни, до комплексни алгоритми от сферата на машинното учене като невронни мрежи и рандомизирани класификационни гори [1], [2]. Подобна богата палитра от прогностични методи по естествен начин поставя въпроса за оптималния такъв, като това е тествано в редица емпирични изследвания. Едно от първите такива е работата на Макридакис и др. [3] през втората половина на седемдесетте години на двадесети век, която има за цел да сравни резултатите от различни прогностични методи върху 111 различни набори от данни (т. нар. състезание М).

Тези изследвания са продължени и разширени в последващи разработки, като в началото на осемдесетте години на миналия век Макридакис и др. [4] представят резултати от състезанието М-2, при което се разглеждат резултатите от различни подходи върху много по-голяма извадка от 1001 набора от времеви редове. Последното състезание на прогностични методи – М3 – приключва през 2000 г. и разглежда още по-голям набор от 3003 различни времеви редове и 24 прогностични метода. Основните изводи от тази изследователска програма, както и от други емпирични разработки, са че макар най-сложните методи не винаги да водят до най-добрите резултати, то ползата от прилагането на формални статистически модели е значителна в сравнение с използването на наивни прогнози. Тези „състезания“ показват и че модели на времеви редове от типа ARIMA има сравнително висока прогностична точност. Това важи с особена сила в случаите на макроикономически и финансови данни. В настоящата разработка показваме как ARIMA моделите могат да бъдат автоматизирани, така че да се достигне до качествени прогнози на изключително ниска цена.

Основната цел на този клас модели е да използва информацията, която вече се съдържа във времевия ред и по този начин да моделира настоящата реализация на дадена променлива като функция от нейните минали реализации. В най-простата версия на модела, настоящата стойност на определена променлива y_t е представена като среднопретеглена сума от p минали нейни стойности y_{t-p} и грешка (ε_t). Този AR(p) модел се дефинира както следва:

$$y_t = \theta + \sum_{i=1}^{p} \beta_i y_i + \varepsilon_t \quad (1)$$

Допълнителна информация може да се съдържа и в самата структура на грешките на модела. Тази информация се използва като се добави плъзгаща се средна на

остатъците. При използване на q минали стойности на грешката, имаме MA(q) модел от следния вид:

(2) $$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \alpha_i \epsilon_i$$

Комбинирайки двете уравнения (1) и (2) получаваме класически модел от типа ARMA(p, q) :

(3) $$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_i + \sum_{i=1}^{q} \alpha_i \epsilon_i + \epsilon_t$$

Ако времевият ред е интегриран от определен порядък d, това също следва да се вземе под внимание като се достига в крайна сметка до модел от типа ARIMA(p, d, q). Като силна страна на този клас модели се откроява възможността за включване и на сезонен компонент, който отразява представената дотук структура на несезонните такива. Включването му води до модел от типа ARIMA(p, d, q)(P, D, Q), който отразява авторегресионния компонент, компонента на плъзгащата се средна, както и потенциалната интеграция в сезонния компонент на разглежданите данни. Повече детайли за този тип модели може да се намери в работата на Хамилтън [6].

Наличието на удобни инструменти за анализ на времеви редове предпоставя и интереса към автоматичните прогнозиране с подобни методи. Ранните разработки могат да бъдат отнесени към деветдесетте години на миналия век, като състезанието М-3 допълнително засилва интереса към тях. Изследване на Мелард и Пастеелс [7] показва базисна софтуерна имплементация на автоматично прогнозиране с ARIMA методи, която изчислява осем модели и избира най-добрия от тях на база най-ниска автокорелация на реда от грешки. Въпреки че този подход има своите ограничения и подбора на оптимален модел има нужда от значително подобрение, тази разработка ясно демонстрира възможностите за автоматизирана аналитика върху времеви редове. Подобряване на прецизността при идентифициране на особеностите на времевите редове е от особена важност и е обект на множество последващи изследвания като например работа на Адиа и др. [8]. Някои изследвания се фокусират и върху алтернативни методи за широкомащабно автоматично прогнозиране [9], но техният ефект върху теорията и практиката е все още ограничен.

Хайндман и Кандакар [10] разработват алгоритъм за автоматично прогнозиране, изпълнен в езика за статистическо програмиране R. Той може да изчислява както модели на експоненциално изглаждане, така и да използва ARIMA методи, което предоставя редица възможности при анализа. От особена важност тук е изборът на оптимален модел въз основа на предварително зададен информационен критерий. Подобен метод на селекция е предпочитан, тъй като води то ясно предпочитание към един определен модел сред широк набор от потенциални алтернативи. Това позволява напълно автоматично прогнозиране след като са взети няколко първоначални ключови решения за действието на алгоритъма. В следващия раздел описваме тези решения, обосноваваме избора на параметри и тестваме подхода върху макроикономически данни за българската икономика.

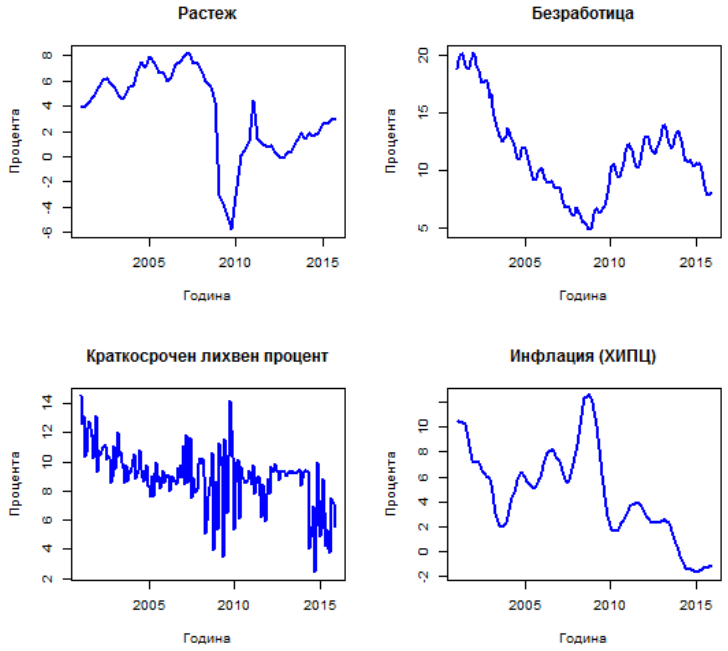
III. Автоматично прогнозиране на времеви редове

Четири основни макроикономически променливи представляват особен интерес за бизнес аналитика и процеса на формиране на публични политики. Това са динамиката на БВП (реалният растеж), който отразява разполагаемите доходи и нивата на

икономическо развитие; нивото на безработица, което е индикатор за наличието на свободен работен ресурс и за цената на труда; лихвените проценти като цена на капитала; и инфлацията, която отразява ценовата динамика и е индикатор за икономическата стабилност. Тяхното навременно и коректно прогнозиране е от ключова важност за планирането и успешните интервенции както на фирмено, така и на национално ниво. Тук използваме тези променливи като примери, с които конструираме автоматична аналитика.

Източникът на данните е Евростат, както и Българската народна банка, а времевите редове покриват периода от 2001 до 2015 г. Времевият ред на растежа е тримесечен, докато безработица, инфлацията (среден ХИПЦ за предходните 12 месеца) и лихвите за краткосрочни бизнес кредити (под 1 г.) са месечни. Тяхната динамика може да бъде проследена на фиг. 1. До икономическата криза, започнала през 2009 г., българската икономика расте устойчиво, като отбелязва както спад на безработицата, така и на лихвените проценти. Този период се характеризира и с растяща инфлация, която понякога достига двуцифрени стойности. Подобна динамика будеше притеснения за потенциалното прегряване на националната икономика.

Икономическата ситуация се промени драстично със започването на глобалната икономическа и финансова криза. През 2009-2010 г. темпът на растеж на България се срина, а безработицата нарасна значително. Усещането за увеличен стопански риск и икономическата несигурност доведе до увеличаване на колебанията на краткосрочните лихвени проценти. Паралелно с това се наблюдаваше и намаляване на инфлационния натиск. Към края на разглеждания период през 2014-2015 г. икономиката до голяма степен се възстанови и растежът бележи известно ускорение. Това води до намаление на безработицата и лек спад на лихвените проценти за бизнеса. Тяхната волатилност остава значителна, като това е съпровождано и с известен дефлационен натиск.



Фигура 1: *Динамика на основни макроикономически променливи, източник: Евростат, БНБ*

Накратко, наблюдаваме значителни колебания на времевите редове, промяна на режима на действие на икономическата система и клъстери на волатилност в някои от променливите (лихвени проценти). Автоматичното моделиране на подобни времеви редове представлява значително предизвикателство, но и разкрива редица възможности. За целта може да се използват модели от класа ARIMA. Основен въпрос при конструирането на модели от този клас е броят лагове (минали реализации), с помощта на които ще моделираме всяка променлива. Това решение зависи както от периодичността им, така и от теоретичните и емпирични очаквания за силата на влияние и инерцията на показателите. Като абсолютен минимум следва да включим лагове, които отразяват поне една година от историята на променливата, а ако тя е поволатилна – поне две години. По подобен начин следва да подходим и със сезонността. Отбелязваме, че е сравнително необичайно икономическите и бизнес времеви редове да бъдат интегрирани от порядък по-висок от втори или трети и залагаме този максимум.

Таблица 1: Параметризация на ARIMA моделите за различните променливи

	p	**d**	**q**	**P**	**D**	**Q**	**No.**
Растеж	8	3	8	4	2	4	19,400
Безработица	12	2	12	2	1	2	9,126
Лихвен процент	12	2	12	2	1	2	9,126
Инфлация	12	2	12	2	1	2	9,126

Параметризацията на моделите за всеки от разглежданите времеви редове е представена в Таблица 1. При така избраната лагова структура достигаме до алгоритъм, който изчислява 19,400 модела за растежа и 9,126 модела за всяка от другите променливи и избира най-оптималният от тях за прогностични цели. За да тестваме коректно прогностичната точност, разделяме данните в две извадки. Първата от тях – обучителната извадка – обхваща периода от 2001 до 2014 и се използва за изчисляване на моделите. Данните от 2015 г. са тестовите данни, с помощта на които се измерва точността на прогнозите извън обучителната извадка. По този начин можем да сравним прогнозите за променливите с техните реализации за 2015 г. и да оценим качествата на предложените модели.

Първоначално алгоритъмът изчислява всички алтернативи модели, използвайки обучителната извадка. Следващата стъпка е изборът на оптималните четири модела от изчислените 46,818 алтернативи. За целта може да се използва някакъв вид информационен критерий. В практиката са популярни три критерия: Бейсовия информационен критерий (BIC), информационния критерий на Акайке (AIC), както и коригирания информационен критерий на Акайке (AICc) [11]. За да формализираме идеята за качество на даден модел първоначално дефинираме функция на правдоподобност L, която отразява вероятността p да наблюдаваме данните x при даден модел M с набор от параметри θ, или:

(4) $$L = p(x|\theta, M)$$

Означавайки максимизираната стойност на тази функция с L_{max}, то дефинираме Бейсовия критерий BIC за модел с k параметри по следния начин:

(5) $$BIC = -2\ln L_{max} + k\ln n$$

В своята същност информационният критерий е мярка за качество на модела, която представя размерът на информационната загуба когато данните се представят от дадения модел. Затова и той може да послужи за избор на най-добрия модел между набор от алтернативи, като се вземе под внимание изборът между обяснителна сила и компактност на модела (брой параметри). За определен набор от данни по-добрите модели бележат по-ниски стойности на техните информационни критерии. Бейсовия критерий е често критикуван поради своята невъзможност да се справи с комплексни набори от модели или параметри, като е валиден при условия, че броят наблюдение е много по-висок от броя използвани параметри ($n >> k$). Това, заедно с други съображения за коректността на неговите резултати, води до растящ консенсус в литература, че е по-удачно използването на критерия на Акайке (AIC) [12]. Той се дефинира по следния начин:

(6) $$AIC = 2k - 2\ln L_{max}$$

Оценката на AIC е асимптотично валидна, което и означава, че следва да се направят корекции за извадки с краен размер, което и води до коригираната версия на този критерий (AICc). Формулата за времеви редове с нормално разпределени грешки е както следва:

(7) $$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

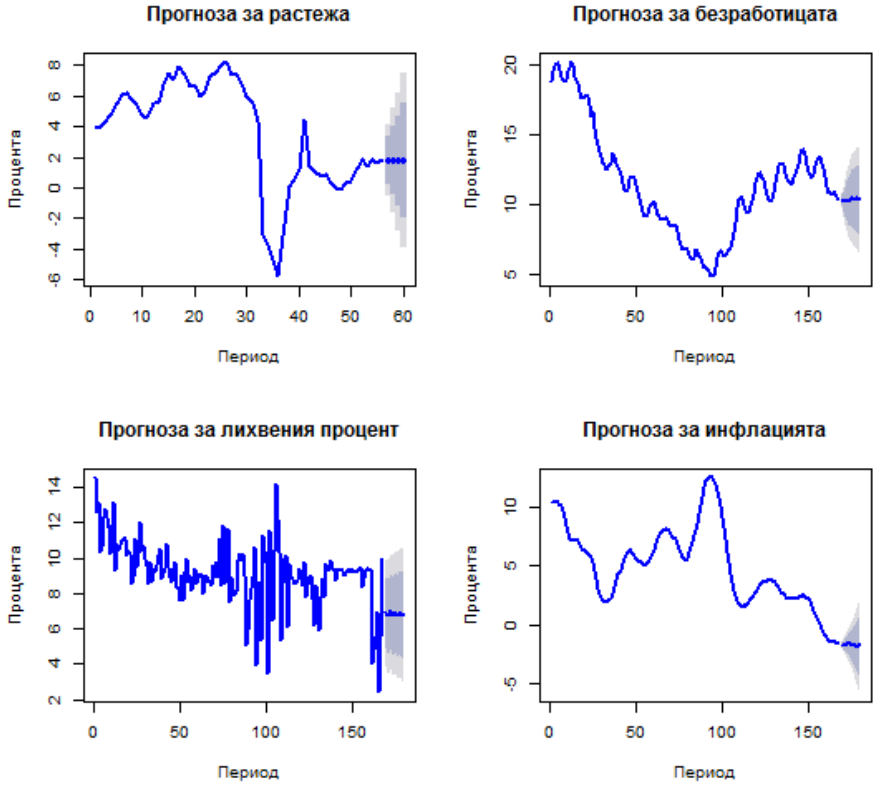
Този критерий е по-чувствителен към модели с повече параметри спрямо некоригираната версия на критерия на Акайке и в този смисъл ще доведе до избора на по-компактни модели. Все пак следва да се има предвид, че с нарастването на размера на извадката коригираната версия на критерия ще конвергира към стойностите на некоригираната такава и поради тази причини се препоръчва използването на AICc във всички случаи [11]. Следвайки литературата, тук също използваме него за избор на оптимален модел за разглежданите времеви редове, но също така представяме и стойностите на алтернативните критерии. От емпирична гледна точка разликата между числените стойности между критериите е много малка и използването на който и да било от тях ще доведе до избора на един и същ оптимален модел.

Резултатите от изчисленията са представени в Таблица 2. Прави впечатление, че оптималните модели са сравнително малко по размер като броят лагове е между 3 и 7. Всички разглеждани променливи са най-добре описани от модели с първи разлики (корекция за интеграция от първи ред), което вероятно се дължи на техните ясно изразени трендове. Във всички случаи освен при растежа информацията от структурата на грешките може да бъде ползотворно включена в моделирането.

Таблица 2: *Оптимални модели за разглежданите времеви редове*

	Растеж	**Безработица**	**Лихвен процент**	**Инфлация**
Оптимален модел	ARIMA(5,1,0)	ARIMA(4,1,3)	ARIMA(3,1,3)	ARIMA(7,1,3)
Критерий на Акайке	189.19	30.88	633.64	-175.99
Критерий на Акайке, коригиран	190.94	31.79	634.35	-174.29
Бейсов критерий	201.24	55.82	655.47	-141.69

Използваме така избраните оптимални модели, за да генерираме прогнози за период от една година. В случая на БВП това означава да се генерира прогноза за четири период напред, а за останалите променливи – за 12 периода напред. Тези прогнози са представени графично на фиг. 2, където ясно можем да проследим разширяващите се доверителни интервали с отдалечаването на прогнозата от реално наблюдаваните данни, както и типичното изглаждане на тренда. Формално тестваме качеството на представените оценки, като ги сравняваме с реализациите на променливите и представяме резултатите в Таблица 2. В нея се съдържат оценките на средната грешка (ME), корен от средната квадратична грешка (RMSE), средната абсолютна грешка (MAE), средната процентна грешка (MPE), средната абсолютна процентна грешка (MAPE), както и средната абсолютна скалирана грешка (MASE).



Фигура 2: *Автоматични прогнози за разглежданите променливи за период от една година*

Вземайки под внимание значителните колебания на времевите редове през разглеждания период, резултатите са повече от добри. Средната абсолютна грешка за всички времеви редове е 0.49 за обучителната извадка и 1.06 за тестовата извадка. Това означава, че автоматизираната процедура произвежда оценки, които имат средно отклонение от 1 процентен пункт от действителните реализации на променливите. Тези резултати се различават в зависимост от разглежданите променливи, като лихвените проценти имат най-големи средни абсолютни грешки и средни абсолютни процентни грешки. Безработицата и инфлацията изглежда са по-лесни за прогнозиране, като техните средни абсолютни грешки са съответно 1.38 и 0.32. В случая на

инфлацията отклонението е особено малко, като подобна точна прогноза може да има голяма практическа стойности.

Времевият ред на растежа също е добре описан от автоматизирания модел – средната абсолютна грешка в обучителната извадка е 0.64, а в тестовата – 1.01. Това означава, че очакваме средно отклонение на прогнозата от реализацията в порядъка на един процентен пункт. Взимайки под внимание комплексните определители на растежа, както и значителните влияние на различни екзогенни шокове, подобни резултати са определен успех. От особено значение за динамиката на БВП през 2015 г. е голямото количество разплатени средства по линия на Европейския съюз поради приключването на програмния период 2007-2013. Резкият ръст на публичните инвестиции оказа своето влияние върху растежа и други макро-агрегати и в този смисъл реализираните отклонения на автоматичната прогноза от реализацията са изненадващо малки.

Table 2: *Metrics for Forecast Accuracy of Optimal Models*

	ME	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**
Растеж						
Обучителни данни	-0.05	1.19	0.64	ND	ND	0.78
Тестови данни	1.01	1.02	1.01	35.91	35.91	1.23
Безработица						
Обучителни данни	-0.03	0.25	0.18	-0.17	1.67	0.53
Тестови данни	-1.17	1.70	1.38	-14.72	16.70	4.12
Лихвен процент						
Обучителни данни	-0.19	1.53	1.05	-5.92	14.20	0.80
Тестови данни	-0.94	1.80	1.55	-23.97	31.50	1.18
Инфлация						
Обучителни данни	-0.01	0.13	0.10	-0.14	3.21	0.37
Тестови данни	0.31	0.39	0.32	-25.08	25.66	1.15
Средно						
Обучителни данни	**-0.07**	**0.77**	**0.49**	**-2.08**	**6.36**	**0.62**
Тестови данни	**-0.20**	**1.23**	**1.06**	**-6.97**	**27.44**	**1.92**

Представените тук примери могат да се разглеждат като индикация за потенциала на автоматизираната аналитика и особено прогнозиране. Показателите за точност на прогнозите показват, че подобни модели се справят сравнително успешно дори при предизвикателни реалистични ситуации. Вероятно е при преминаването от макроикономически към микроикономически променливи (пр. продажби), които са по-слабо чувствителни към екзогенни шокове и имат ясно изразен тренд, точността на прогнозата да се увеличи допълнително. Също така следва да имаме предвид, че сравнението на точността на генерираните прогнози би било напълно коректно само при тяхното сравнение с реалистична алтернатива (т.е. с други прогнози), при което разликите между стойностите вероятно ще бъдат дори още по-малки.

IV. Заключение

Автоматизираната аналитика предлага огромни възможности в един свят на увеличаващи се масиви от данни и растящ интерес към анализа, който обаче е ограничен от обективно разполагаемите ресурси за това. Настоящата разработка разгледа възможността за изчисление на голям набор от алтернативни ARIMA модели и избо-

ра на оптималния такъв с цел генериране на автоматични прогнози на времеви редове. Изследването на четири макроикономически променливи показа, че този подход генерира много добри прогнози дори в условия на подчертана динамика и неопределеност на средата.

Тези резултати откриват и редица възможности за бъдещи изследвания. От особен интерес би било да се разгледат алтернативни алгоритми, които също подлежат на автоматизация и да е оцени тяхната прогностична точност. Потенциални такива биха били представители от семейството на екпоненциалните плъзгащи се средни или алгоритми за машинно научаване като рандомизираната класификационна гора. Друг важен въпрос е относно относителната точност на прогнозите в зависимост от типа данни, тяхната честота и произход. Не на последно място, автоматизираната аналитика може да бъде приложена и в други области от икономиката и бизнеса.

В заключение, автоматизираната аналитика разкрива възможности за прехвърлянето на типични задачи от анализатора към определени алгоритми, като по този начин се освобождава организационен ресурс, който може да се съсредоточи върху по-сложните проблеми на стратегирането, планирането и управлението на днешната комплексна икономическа и бизнес среда.

Литература:

1. Hyndman, Rob J., and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.
2. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
3. Makridakis, Spyros, Michele Hibon, and Claus Moser. „Accuracy of forecasting: An empirical investigation." *Journal of the Royal Statistical Society. Series A (General)* (1979): 97-145.
4. Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. „The accuracy of extrapolation (time series) methods: Results of a forecasting competition." *Journal of forecasting* 1, no. 2 (1982): 111-153.
5. Makridakis, Spyros, and Michele Hibon. „The M3-Competition: results, conclusions and implications." *International journal of forecasting* 16.4 (2000): 451-476.
6. Hamilton, James Douglas. *Time series analysis*. Vol. 2. Princeton: Princeton university press, 1994.
7. Melard, Guy, and J-M. Pasteels. „Automatic ARIMA modeling including interventions, using time series expert software." *International Journal of Forecasting* 16.4 (2000): 497-508.
8. Adya, Monica, Fred Collopy, J. Scott Armstrong, and Miles Kennedy. "Automatic identification of time series features for rule-based forecasting."*International Journal of Forecasting* 17, no. 2 (2001): 143-157.
9. Chakrabarti, Deepayan, and Christos Faloutsos. „F4: large-scale automated forecasting using fractals." *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002.
10. Hyndman, R. J., and Y. Khandakar. „Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software 27.3* (2008): 1-22.
11. Burnham, Kenneth P., and David R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.
12. Burnham, Kenneth P., and David R. Anderson. „Multimodel inference understanding AIC and BIC in model selection." *Sociological methods & research* 33.2 (2004): 261-304.