МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ОБЧИСЛЮВАЛЬНІ МЕТОДИ

УДК 004.4, 004.6

Т. А. Чупілко, Ю. В. Ульяновська, М. Ф. Мормуль, А. Е. Лагода

PYTHON ДЛЯ ОБРОБКИ ДАНИХ І МОДЕЛЮВАННЯ ФІНАНСОВО-ЕКОНОМІЧНИХ ПОКАЗНИКІВ

Університет митної справи та фінансів, Дніпро

Анотація. У статті розглянуто аспекти ефективної обробки даних. Значна увага приділяється проблемам, що виникають при моделюванні і прогнозуванні даних та роль досліджень для прийняття рішень. Визначаються етапи роботи з даними та особливості, що є притаманними кожному етапу. Особливе місце у роботі займає опис можливостей програмної обробки даних з використанням мови Python, яка набуває все більшої популярності завдяки простоті, гнучкості, відкритому коду, зручності роботи з даними у різних форматах, а також багатьом розробленим пакетам, які сприяють швидкій та ефективній обробці інформації. Розглядаються NumPy, Pandas, які надають структури даних і функції, що дозволяють зробити роботу зі структурованими даними простою і швидкою, найпопулярніший інструмент для візуалізації даних Matplotlib, пакети для різних обчислювальних задач SciPy, Statsmodels, а також пакет, орієнтований на машинне навчання Scikit-learn. Наводиться приклад використання Python для задач митної сфери. Авторами створено програму для розрахунку, в якій використовуються вище зазначені пакети. Будуються регресійні моделі для аналізу поповнення державного бюджету України надходженнями від митних органів за рахунок ввізного та вивізного мита. Проводиться аналіз моделей на основі економетричних методів моделювання та розраховуються прогнозні оцінки надходжень.

Ключові слова: Python, обробка даних, моделювання, прогнозування, регресійна модель.

Аннотация. В статье рассмотрены аспекты эффективной обработки данных. Значительное внимание уделяется проблемам, возникающим при моделировании и прогнозировании данных, а также роль исследований для принятия решений. Определяются этапы работы с данными и особенности, которые присущи каждому этапу. Особое место в работе занимает описание возможностей программной обработки данных с использованием языка Python, которая приобретает все большую популярность благодаря простоте, гибкости, открытому коду, удобству работы с данными в различных форматах, а также многим разработанным пакетам, которые способствуют быстрой и эффективной обработке информации. Рассматриваются NumPy, Pandas, которые создают структуры данных и функции, позволяющие сделать работу со структурированными данными простой и быстрой, самый популярный инструмент для визуализации данных Matplotlib, пакеты для различных вычислительных задач SciPy, Statsmodels, а также пакет, ориентированный на машинное обучение Scikit-learn. Приводится пример использования Python для задач таможенной сферы. Авторами создана программа для расчета, в которой используются вышеуказанные пакеты. Строятся регрессионные модели для анализа пополнения государственного бюджета Украины поступлениями от таможенных органов за счет ввозной и вывозной пошлины. Проводится анализ моделей на основе эконометрических методов моделирования и рассчитываются прогнозные оценки поступлений.

Ключевые слова: Python, обработка данных, моделирование, прогнозирование, регрессионная модель.

Abstract. The article considers aspects of efficient data processing. Much attention is paid to the problems that arise in modeling and forecasting data and the role of research in decision-making. The stages of work with data and features that are inherent in each stage are determined. A special place in the work is described by the possibilities of software data processing using Python, which is becoming increasingly popular due to simplicity, flexibility, open source, ease of working with data in various formats, as well as many developed packages that facilitate fast and efficient information processing. NumPy, Pandas, which provide data structures and functions that make working with structured data simple and fast, the most popular tool for data visualization Matplotlib, packages for various computational tasks SciPy, Statsmodels, as well as a package focused on machine learning Scikit-learn. An example of using Python for customs tasks is given. The authors have created a program for calculation, which uses the above packages. Regression models are being built to analyze the replenishment of the state budget of Ukraine with revenues from customs authorities at the expense of import and export duties. The analysis of models on the basis of econometric methods of modeling is carried out and forecast estimates of receipts are calculated.

Key words: Python, data processing, modeling, forecasting, regression model.
DOI: https://doi.org/10.31649/1999-9941-2021-51-2-68-77.

Вступ

В сучасному світі компанії, підприємства, установи мають справу з великим обсягом даних. Останнім часом все більшої популярності набувають технології, що дозволяють працювати з великими масивами інформації. Підхід до обробки даних залежить, насамперед, від їх типу, мети використання, можливостей підприємства чи установи в організації збору, систематизації, аналізу даних, які часто бувають обмеженими. Наразі невелика кількість підприємств і установ мають змогу інвестувати значні кошти в розвиток цих технологій для власних потреб. Великі компанії, безумовно, мають можливості для розвитку бізнес-аналітики. Для більшості технік роботи з даними обсяг даних може бути і великим, і малим. Наскільки великі дані потрібні для результату, визначається інтересами компанії. Не завжди компанії мають справу із генерованими даними великих обсягів. Найчастіше є певна база даних компанії, яку і потрібно використовувати.

До основних технік аналізу відносяться кластерний та факторний аналіз, моделювання і прогнозування на основі економетричних та оптимізаційних методів, визначення викидів, штучний інтелект, сітьові графи, машинне навчання. Деякі з цих методів розвивалися давно, інші з'явилися нещодавно. При правильному виборі різні техніки і технології є однаково ефективними для оцінювання ситуації в різних сферах діяльності та прийняттю відповідних управлінських рішень. Тому важливим є розуміння, яка те-

хніка для розв'язання якої проблеми підходить, як ці техніки працюють, і як їх застосовувати для моделювання показників.

В різних технологіях для вивчення даних використовуються схожі базові математичні інструменти: і у стандартних пакетах обробки даних та відповідних бібліотеках і модулях у популярній об'єктно-орієнтованій мові Python, і в широко відомому і розповсюдженому пакеті MS Office, і в інших програмних продуктах на кшталт мови R, що є поширеною для статистичного аналізу даних.

В означеній темі роботи автори використовували роботи іноземних авторів [1], [2], де розглядається використання Python та програмних пакетів саме для аналізу даних. В опублікованих матеріалах [3] − [5] порушені питання, пов'язані з проблематикою цієї роботи. Є певні роботи українських та іноземних вчених щодо міркувань про доступ до великих даних, імплементацію їх у статистику, реальну користь їх і т. ін. Незважаючи на значний науковий доробок щодо загальних питань про дані, дослідження із застосуванням інструментарію ефективної обробки даних із застосуванням мов програмування відсутні, і, зокрема, в митній сфері.

Актуальність

Моделювання охоплює різні галузі, різноманітні показники. Особливо важливим є моделювання фінансово-економічних показників, яке дає змогу прогнозувати їх значення при наявності певної тенденції, дати оцінку прогнозу точкову та інтервальну, що визначається довірчим інтервалом. Спеціальні критерії дозволяють оцінити якість побудованої моделі. Тож, аналітика даних сприяє покращенню процесу прийняття рішень у будь-якій сфері діяльності. Але, разом з тим, потрібні інструменти для високоефективної та швидкої обробки даних.

Останнім часом мова програмування Python та велика кількість бібліотек з відкритим кодом, які динамічно поповнюються, є дуже популярним і потужним інструментом, що дозволяє ефективно обробляти дані, моделювати і прогнозувати показники, використовуючи сумісну можливість написання коду та готові рішення.

Технології опрацювання даних визначаються характером даних та метою дослідження.

Але є спільні проблеми при використанні різних технологічних інструментів – відбір і підготовка даних, а також фахове опрацювання результатів.

Щоб скористатися навіть абсолютно автоматизованою системою обробки даних, їх потрібно правильно відібрати, відсортувати, нормалізувати, обрати метод аналізу, і після обробки даних програмою провести саме аналіз, інтерпретацію, прогнозування і т. ін.

На етапі підготовки даних виникають певні проблеми, пов'язані з різними форматами даних, отриманих із різних джерел, з обмеженим доступом до даних, обумовленим цінністю даних та конфіденціальністю і суворою регламентованістю. Дані можуть мати різні одиниці виміру, різні рівні агрегування.

Покращення якості даних, розуміння, як дані взаємодіють між собою, оцінювання розподілів і приведення до певного формату неможливе без знання фундаментальних основ відповідного математичного апарату.

Сучасні пакети підтримують багато популярних методів моделювання і оцінювання моделей. Застосування потребує вивчення мови і освоєння пакетів, в яких для використання певного методу чи функції потрібно задати різноманітні параметри для виконання.

Ще одна проблема пов'язана з доступом до даних. Офіційні статистичні дані, які є основою для моделювання і прогнозування фінансово-економічних показників в масштабах країни, зокрема в митній сфері, є досить обмеженими і, в основному, агрегованими. Часто наводяться на офіційних сайтах або в статистичних збірниках дані, що не є нормалізованими., або їх кількість недостатня для побудови адекватних моделей. Тож, задачі моделювання обмежуються тією статистикою, що є офіційно доступною.

Процес моделювання залежить від якості даних і, не менше, від професіоналізму аналітика.

Мета

Метою роботи є аналіз ефективних інструментів для обробки та моделювання даних; застосування Python та бібліотек для аналізу, моделювання і прогнозування фінансово-економічних показників, на прикладі офіційних даних митних надходжень до державного бюджету України по окремим видам,а саме, вивізному та ввізному миту.

Задачі

1. Проаналізувати проблеми у роботі з даними та аспекти застосування інструментів для ефективної обробки даних.
2. Застосувати найбільш придатні інструменти для конкретної задачі моделювання фінансово-економічних показників (на прикладі митних надходжень до бюджету від окремих показників).

Розв'язання задач

Інструменти для ефективної та швидкої обробки даних

Для обробки даних доцільно використовувати технології, найбільш прийнятні для певного типу і задач, що розв'язує дослідник.

У деяких задачах достатньо використання зручного, зрозумілого і доступного інструменту, як MS Excel, що має широкі можливості і пакет аналізу, хоча і дещо обмежений, але придатний для отримання результатів у першому наближенні для уявлення про характер даних. За допомогою електронних таблиць неможливо в програмному режимі запустити виробничу модель, наприклад, штучного інтелекту, але за їх допомогою можна проаналізувати характер даних, змоделювати і спрогнозувати результат. Цей результат можна отримати на основі класичних підходів теорії ймовірностей та математичної статистики щодо нормування даних, кореляційного та регресійного аналізу, оцінювання прогнозних точкових та інтервальних значень, а також за допомогою процедур для визначення оптимальних розв'язків лінійних та нелінійних задач оптимізації.

Для використання автоматизації обробки даних в програмному режимі потрібні знання тієї чи іншої мови програмування. Але для розуміння сутності аналізу, що використовується в технологіях обробки і різних прикладних пакетах таких, як Statistika, SPSS і т. ін., не обов'язково знати, як пишеться код. Ці потужні інструменти включають різноманітний аналіз, в тому числі, регресійний, факторний, кластерний, побудову моделей за допомогою нейронних мереж і багато іншого, а також дають можливість отримати графічне відображення результатів, якщо дозволяє розмірність і постановка задачі.

У процесі роботи з даними можна виділити декілька етапів.

1. Призначення цілі дослідження. При цьому готується проектне завдання і оцінюються мета дослідження і вартість роботи.
2. Збір і підготовка даних, так званий «розвідувальний аналіз». Певні складнощі виникають уже на цьому етапі. Дані можутьбути розрізненими, в різних форматах і потребують нормалізації і приведення до однородності. Матриці можуть бути не повністю заповненими, виродженими. Обов'язково потрібно підібрати алгоритм для заповнення порожнеч. Нерідко в даних бувають значні відхилення, тобто, викиди, які потрібно усунути, тобто очистити дані. Інакше ніякі методи моделювання не приведуть до адекватної моделі. Таким чином, процес підготовки даних є дуже кропітким і рутинним, майже «ручним», потребує інтелектуального підходу і розуміння цілі дослідження.
3. Аналіз і моделювання даних, тобто вибір моделі та оцінювання її параметрів, в машинному навчанні – «тренування моделі». На цьому етапі потрібно зрозуміти, як дані пов'язані між собою, оцінити розподіли даних, визначити і усунути викиди, а також перевірити наявність мультиколінеарності в системі та таких негативних явищ, як гетероскедастичність та автокореляція, які потребують додаткових перетворень змінних та особливих методів. Для цього використовуються певні статистичні методи і просте моделювання. Постають питання: чи пов'язані між собою досліджувані фактори і показник, чи є мультиколінеарність в системі даних, чи можна зменшити кількість змінних і тим самим спростити модель, яку форму залежності обрати для моделювання, яким способом звести модель до лінійної форми і т. ін. На цьому етапі потрібні знання предметної області, а також математики, теорії ймовірностей і математичної статистики. Тільки після указаних досліджень і перетворень даних можна скористатися готовими рішеннями − пакетами програм. Сам процес моделювання − «тренування моделі» − означає побудову різних моделей на одному наборі даних, випадково відібраних із загальної сукупності. Кількість даних можна варіювати з допомогою параметрів, що задаються для обраного методу. Можна обрати найкращоу модель за певними критеріями, наприклад, метод найменших квадратів, або метод, заснований на дереві рішень (знову ж таки, з різними параметрами, що можна варіювати), або метод абсолютних відхилень і т.п. Можна тренувати набір даних декілька разів, змінюючи параметри, і таким чином, досягти найкращого результату. Тож, побудова моделі – ітераційний процес і потребує навичок дослідника.
4. Перевірка адекватності моделі і значимості факторів моделі. Після отримання найкращого результату (наприклад, порівнюються сума квадратів відхилень і обирається набір параметрів, що дає найменше із усіх) оцінювання якості моделі відбувається за статистичними критеріями. Якщо якість незадовільна, то мдель потрібно «перенавчити».
5. Застосування моделі до незнайомих даних − так званий «тренувальний сет» обирається із тієї ж вибірки − «прогностичне моделювання», тобто визначається прогноз.

Описаний підхід застосовується для задач моделювання і прогнозування при машинному навчанні(Machine Learning). Python, наприклад, має свою бібліотеку Scikit-learn з різноманітними алгоритмами.

Машинне навчання наразі є дуже популярною і перспективною технологією серед аналітиків (data-scientists). Ринок машинного навчання швидко зростає. З 2016 року його обсяг подолав позначку в $ 1

млрд, а до 2025 року, судячи з прогнозів, він може збільшитися до $ 39,98 млрд. 60% компаній в світі вже використовують машинне навчання.

Серед завдань, які можуть вирішуватися засобами машинного навчання, можна зазначити задачі моделювання і прогнозування показників в залежності від одного або декількох факторів або оптимізаційні задачі. Використовуються як традиційні методи економетричного аналізу, включаючи однофакторні, багатофакторні моделі на основі методу найменших квадратів, так і нетрадиційні, типу, дерева рішень з великою кількістю встановлюваних параметрів, що дають гнучкість моделювання параметрів моделей.

Набувають популярності так звані «нейронні мережі». При моделюванні використовуються поняття ризику, кількісні ознаки якого обчислюються у відповідності до числових характеристик дискретних та неперервних випадкових величин.

За останні десять років Python перетворився в одну із найважливіших мов програмування, застосовуваних у науці про дані, в машинному навчанні та розробці програмного забезпечення загального призначення в академічних установах і промисловості. Поліпшені бібліотеки для Python спияли тому, що він став серйозним конкурентом в рішенні задач створення додатків обробки даних.

У багатьох сучасних середовищах застосовується загальний набір успадкованих бібліотек, написаних на FORTRAN і C, що містять реалізації алгоритмів линійної алгебри, оптимізації, інтегрування та ін. Тому численні компанії використовують Python як «клей» для об'єднання написаних за багато років програм.

Пакети Python для роботи з даними

NumPy, скорочення від «Numerical Python», − основний пакет для виконання наукових розрахунків на Python. Поверх NumPy побудовано інші бібліотеки.

Основні можливості пакету: швидко і ефективно можна створювати об'єкти багатовимірних масивів ndarray; має функції для виконання обчислень з елементами одного масиву або математичних операцій з декількома масивами; надає засоби для читання і запису на диски наборів даних, представлених у вигляді масивів; викоритстовує операції лінійної алгебри, перетворення Фур'є і генератор випадкових чисел; має засоби для інтеграції з кодом, що написаний на C, C++ або Fortran.

NumPy значно прискорює роботу з масивами. Як засіб зберігання і маніпуляції даними, масиви NumPy значно ефективніші за вбудовані в Python структури даних.

Багато засобів обчислень, орієнтовані на Python, або використовують масиви NumPy в якості основної структури даних, або якимось іншим способом організують інтеграцію з NumPy.

Pandas надає структури даних і функції, що дозволяють зробити роботу зі структурованими даними простою і швидкою. Завдяки цій бібліотеці Python перетворився в потужне і продуктивне середовище аналізу даних. Основні об'єкти pandas − це *DataFrame* − двовимірна таблиця, в якій рядки і стовпці мають мітки, і *Series* − об'єкт одновимірного масиву з мітками.

У бібліотеці pandas поєднуються висока продуктивність засобів роботи з масивами, притаманна NumPy, і гнучкі можливості маніпулювання даними, властиві електронним таблицям і реляційним базам даних (наприклад, на основі SQL). Оскільки маніпулювання даними, їх підготовка і очищення грають дуже велику роль в аналізі даних, pandas є одним з основних інструментів.

Основні можливості бібліотеки: має розвинені засоби індексування, що дозволяють просто змінювати форму наборів даних, формувати зрізи, виконувати агрегування і вибирати підмножини; структури даних з позначеними осями підтримують автоматичне або явне вирівнювання даних, що виключає появу типових помилок при роботі з невирівняні даними і даними з різних джерел, які по-різному індексовані; має вбудовану функціональність часових рядів; одні і ті ж структури даних здатні підтримувати як часові ряди, так і дані інших видів; арифметичні операції, чкі виконуються з об'єктами, як з числовими даними; має гнучку обробку відсутніх даних (дозаповнення); інтеграція даних; підтримка з'єднання і інших реляційних операцій, наявних в популярних базах даних (наприклад, на основі SQL).

Багато засобів, присутні в pandas, або є частиною мови R, або надаються додатковими пакетами.

Сама назва pandas утворена від *panel data* (панельні дані), що застосовуються в економетриці для позначення багатовимірних структурованих наборів даних, так і від фрази Python data analysis.

Matplotlib − найпопулярніший в Python інструмент для створення графіків і інших способів візуалізації *двовимірних* даних, підходить для створення графіків, придатних для публікації. Хоча є можливості візуалізації в інших пакетах, matplotlib використовується найчастіше і тому добре інтегрована з іншими частинами екосистеми.

SciPy − набір пакетів, призначених для вирішення різних стандартних обчислювальних задач. Деякі з них: scipy.integrate − підпрограми чисельного інтегрування і розв'язання диференціальних рівнянь; scipy.linalg − підпрограми лінійної алгебри і розкладання матриць, доповнюють ті, що включені в numpy.linalg; scipy.optimize − алгоритми оптимізації функцій (знаходження екстремумів) і пошуку коренів; scipy.signal − засоби обробки сигналів; scipy.sparse − алгоритми роботи з розрідженими матрицями і розв'язання розріджених систем лінійних рівнянь; scipy.special − обгортка навколо

SPECFUN, написаної на Fortran-бібліотеці, що містить реалізації багатьох стандартних математичних функцій, в тому числі гамма-функції; scipy.stats − стандартні безперервні і дискретні розподіли ймовірностей (функції щільності ймовірності, формування вибірки, функції безперервного розподілу ймовірності), різні статистичні критерії і додаткові описові статистики.

Scikit-learn є основним інструментарієм програмістів для машинного навчання на Python. У ньому є підмодулі для наступних моделей: класифікація: метод опорних векторів, метод найближчих сусідів, випадкові ліси, логістична регресія і т. ін.; регресія: Lasso, гребневая регресія і т. ін.; кластеризація: метод k середніх, спектральна кластеризація і т. ін.; зниження розмірності: метод головних компонент, відбір ознак, матрична факторизація і т. ін .; вибір моделі: пошук на сітці, перехресний контроль, метрики; попередня обробка: виділення ознак, нормування.

Scikit-learn орієнтований головним чином на прогнозування і передбачення.

Statsmodels − пакет статистичного аналізу. У порівнянні із Scikit-learn, пакет Statsmodels містить алгоритми класичної (перш за все частотної) статистики та економетрики. У нього входять наступні підмодулі: регресійні моделі: лінійна регресія, узагальнені лінійні моделі, лінійні моделі зі змішаними ефектами і т. ін.; дисперсійний аналіз (ANOVA); аналіз часових рядів: AR, ARMA, ARIMA, VAR і інші моделі; непараметричні методи: ядерна оцінка щільності, ядерна регресія; візуалізація результатів статистичного моделювання.

Пакет statsmodels орієнтований більшою мірою на статистичне виведення, він дає оцінки невизначеності і p-значення параметрів. Використовується разом з NumPy і Pandas.

В Python є бібліотеки для зручного і швидкого зчитування даних в форматах електронних таблиць, баз даних, csv та ін.

Приклад використання Python для моделювання митних надходжень до державного бюджету України

Для цієї роботи були використані дані офіційної статистики [6].

Зазначимо, що інформація у відкритому доступі є дуже обмеженою.

Для задачі є консолідовані дані, які включають повний обсяг надходжень від митних органів до державного бюджету України, а також надходження від ввізного та вивізного мита, наведені в табл. 1.

Таблиця 1 – вихідні дані для моделювання надходжень до державного бюджету України від митних органів всього та за окремими видами.

Рік	Надходження до державного бюджету України, У, грн.	Надходження до державного бюджету України ввізного мита Х1, грн.	Надходження до державного бюджету України вивізного мита, Х2, грн.
2013	1,40036E+11	1,2550E+10	2,4900E+08
2014	3,57084E+11	1,3056E+10	1,7900E+08
2015	5,34694E+11	1,7422E+10	2,4500E+08
2016	6,16219E+11	2,0004E+10	3,7000E+08
2017	6,98405E+11	2,2257E+10	6,4300E+08
2018	8,33615E+11	2,3301E+10	5,1600E+08
2019	8,79833E+11	2,2778E+10	2,3000E+08
2020	8,77603E+11	2,1538E+10	2,5700E+08

Проведемо аналіз окремих складових у надходженнях, зокрема ввізного та вивізного мита.

За даними Таблиці 1 оцінимо наявність і тісноту зв'язку між ними та загальними надходженнями від митних органів, вигляд і тип моделі, параметри регресії, адекватність моделі, статистичну значимість параметрів, наявність автокореляції, визначимо прогнозне значення показника (точкову та інтервальну оцінки) і побудуємо довірчі зони регресії. У якості інструмента моделювання використаємо Python та бібліотеки NumPy, Statsmodels, Matplotlib, Xlrd (для зчитування даних із файлу Excel).

Найпоширенішою сучасною методикою моделювання структурованих даних є економетричне моделювання. За допомогою регресійного аналізу оцінюється залежність показника від одного чи декількох факторів. Найкращий результат дає метод найменших квадратів відхилень вихідних даних від змодельованих. Адекватність моделі оцінимо за допомогою критерію Фішера. Оцінювання статистичної значимості параметрів регресій, а також довірчих інтервалів регресій, проведемо на основі критерію Стьюдента. Окрім цього, отримаємо іншу статистику по моделі і застосуємо модель для прогнозу.

Нижче наведено лістинги результатів програмного виконання розрахунків на Python та графіки для наочного уявлення про побудовані моделі.

Спочатку визначимо, яким чином надходження до бюджету від митних органів залежать від надходжень за ввізне мито. На рис.1 і рис.2 представлені результати виконання.

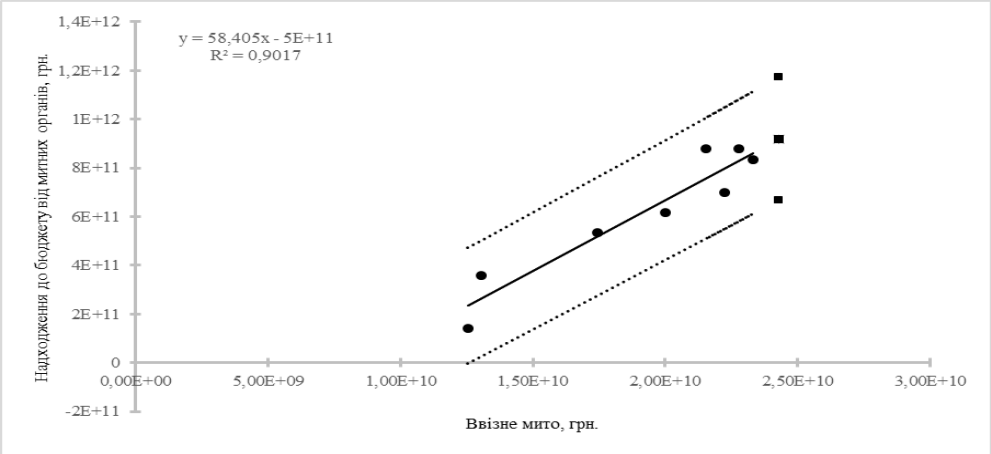

Рисунок 1 – Регресія надходжень до бюджету України від митних органів на ввізне мито, довірчий інтервал прогнозу, довірча зона регресії, побудовані з надійністю 0,95

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                     55.03
Date:                Tue, 03 Aug 2021   Prob (F-statistic):           0.000309
Time:                        22:34:43   Log-Likelihood:                -211.97
No. Observations:                   8   AIC:                             427.9
Df Residuals:                       6   BIC:                             428.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            58.4053      7.873      7.418      0.000      39.141      77.670
const      -4.991e+11   1.54e+11     -3.245      0.018   -8.75e+11   -1.23e+11
==============================================================================
Omnibus:                        1.280   Durbin-Watson:                   1.337
Prob(Omnibus):                  0.527   Jarque-Bera (JB):                0.653
Skew:                           0.145   Prob(JB):                        0.722
Kurtosis:                       1.631   Cond. No.                     9.46e+10
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.46e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
Кореляція:
[[1.     0.9496]
 [0.9496 1.    ]]
Значення фактору для заданого прогнозу показника: [[2.56681055e+10]
 [2.85201172e+10]]

Process finished with exit code 0
```

Рисунок 2 – Лістинг виконання програми (модель залежності митних надходжень до бюджету від ввізного мита)

Проаналізуємо основні результати моделювання.

Застосовано модель OLS, метод Least squares. Залежна змінна – y. Рівняння регресії виведено на рис. 1. Коефіцієнт кореляції дорівнює 0,95, що свідчить про сильну кореляцію між фактором і показником. Коефіцієнт детермінації скорегований − 0,885: зміна показника обумовлена зміною фактору на 88,5%.

F-статистика свідчить про адекватність моделі: розраховане значення дорівнює 55,03, критичне значення для степенів вільності задачі і рівня значимості 0,05 дорівнює 5,98. Розраховані значення t-статистики 7,42 для нахилу та -3,25 для перетину регресії. Обидва параметри є статистично значимими з довірчою імовірністю 0.975. Критичне значення t-статистики 2,45.

Відповідно, довірчі інтервали параметрів регресії при значимості 0,025: для нахилу: (39,14; 77,67), для перетину: (-8,75E+11; -1,23E+11). Статистика Дарбіна–Уотсона свідчить про відсутність автокореляції в моделі. Коваріаційна матриця вірно специфікована.

Модель може бути використана для прогнозу показника. Визначено прогнозні оцінки (точкові та інтервальні). Коефіцієнт еластичності за середніми показниками за останні чотири роки дорівнює 1,69, що означає, що показник є еластичним по фактору, причому темп зростання наповнення бюджету від митних надходжень уповільнюється, зокрема, за рахунок ввізного мита.

Побудуємо і проаналізуємо регресію митних надходжень до бюджету на вивізне мито. Аналогічно до раніше викладеного, отримаємо результат виконання програми. Для наочного представлення результату маємо діаграму з рівнянням регресії на рис.3.

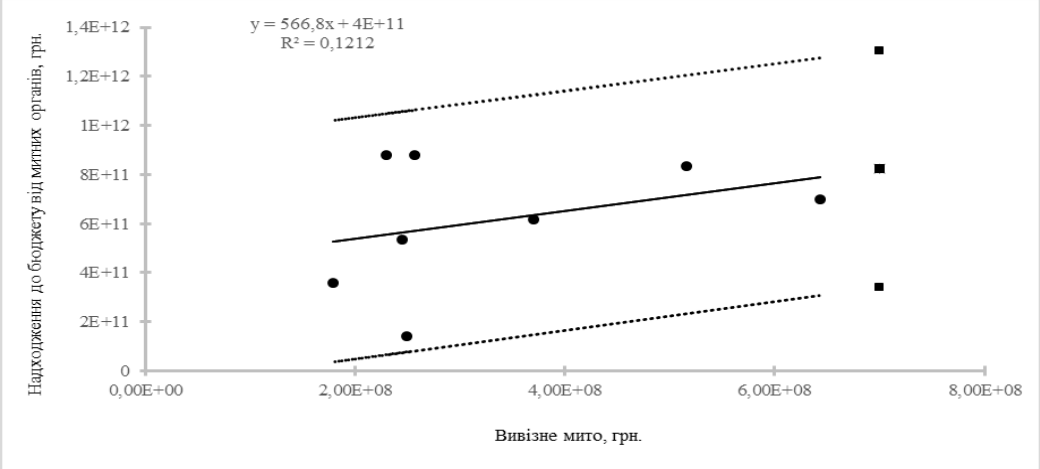

Рисунок 3 – Регресія надходжень до бюджету України від митних органів на вивізне мито, довірчий інтервал прогнозу, довірча зона регресії, побудовані з надійністю 0,95

На рис 4 представлено відповідний лістинг.

Застосовано ту ж модель OLS, що і в попередньому випадку. Залежна змінна – y. Рівняння регресії виведено на рис.5. Коефіцієнт кореляції дорівнює 0,348, що свідчить про дуже слабу кореляцію між фактором і показником. Коефіцієнт детермінації 0,1212 дуже низький, близький до нуля: зміна надходжень від митниці обумовлена зміною надходжень від вивізного мита на 12%. F-статистика свідчить про неадекватність моделі: розраховане значення дорівнює 0,83, критичне значення для степенів вільності задачі і рівня значимості 0,05 дорівнює 5,98. Розраховані значення t-статистики 0,91 для нахилу та 1,86 для перетину регресії. Обидва параметри статистично не значимо відрізняються від нуля з рівнем значимості 0,125. Критичне значення t-статистики 2,45 для степенів вільності моделі і заданого рівня значимості.

Дуже широка довірча зона пояснюється значним розкидом вихідних даних, і, відповідно, широкий довірчий інтервал прогнозу не має практичної цінності. Статистика Дарбіна–Уотсона свідчить про відсутність автокореляції в моделі. Коваріаційна матриця вірно специфікована.

Модель, очевидно, не можна рекомендувати використовувати для прогнозу показника.

Коефіцієнт еластичності за середніми показниками за останні чотири роки дорівнює 0,34, збільшуючись до значення 0,5, тобто показник є нееластичним по фактору, причому темп зростання наповнення бюджету від митних надходжень пришвидчується, за рахунок вивізного мита.

```
Результат розрахунку:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.121
Model:                            OLS   Adj. R-squared:                 -0.025
Method:                 Least Squares   F-statistic:                    0.8276
Date:                Tue, 03 Aug 2021   Prob (F-statistic):              0.398
Time:                        22:29:22   Log-Likelihood:                -220.73
No. Observations:                   8   AIC:                             445.5
Df Residuals:                       6   BIC:                             445.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           566.7962    623.032      0.910      0.398    -957.708    2091.300
const       4.267e+11    2.3e+11      1.856      0.113   -1.36e+11    9.89e+11
==============================================================================
Omnibus:                        0.138   Durbin-Watson:                   0.410
Prob(Omnibus):                  0.933   Jarque-Bera (JB):                0.213
Skew:                          -0.204   Prob(JB):                        0.899
Kurtosis:                       2.313   Cond. No.                     8.94e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
[2] The condition number is large, 8.94e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Кореляція:
[[1.     0.3482]
 [0.3482 1.    ]]
Значення фактору для заданого прогнозу показника: [[4.44254957e+08]
 [4.93616619e+08]]

Process finished with exit code 0
```

Рисунок 4 – Лістинг виконання програми (модель залежності митних надходжень до бюджету від вивізного мита)

Зробимо загальні висновки щодо двох останніх моделей. Для обох моделей використано метод найменших квадратів, що дає найкращу апроксимацію вихідних даних з найменшою сумою квадратів похибок. Оцінки, отримані цим методом за теоремою Гауса-Маркова є ефективними і незміщеними. Для обох моделей розрахунок відбувався за допомогою створеної авторами програми на Python та з використанням бібліотек NumPy, Statsmodels, Matplotlib, Xlrd. Попередня підготовка даних відбувалася в Excel.

В роботі показано можливість програмної обробки даних. Не вдалося отримати більш повні статистичні дані з офіційного сайту, щоб можна було провести аналіз по великих масиивах аналогічних даних. Але проблеми, що виявилися при моделюванні за наведеними агрегованими даними, відтворилися б аналогічно. Окрім того, в лістингах указане застереження, що кількість даних для коректного розрахунку має бути великою. Якщо кількість даних недостатня, то потрібно використовувати відкореговані оцінки, що відомо також з теорії математичної статистики.

Можна зазначити, що при невеликій кількості даних можна було б скористатися статистичним аналізом, що надає Excel, і це було б найефективніше рішення. Але метою роботи було використання можливостей Python для моделювання. Та ж сама програма дала б ефективний розрахунок для великих масивів даних, що не є проблемою в програмуванні безпосередньо. Також можна зазначити, що дані можуть бзберігатися у базах даних. В Python є відповідні засоби роботи і з такими даними.

Тож, маємо відповідні економічні висновки, що виявляють великі проблеми в надходженнях від митних органів і, зокрема, від вивізного мита.

Можна будувати різні моделі однофакторні, і багатофакторні лінійні і нелінійні, і таким чином, отримати повну картину відносно того, які процеси відбуваються в галузі, зокрема, в митній, виявити позитивні і негативні явища. На основі результатів моделювання можна отримати науково обгрунтовані прогнози і приймати відповідні управлінські рішення.

Висновки

1. Проведено аналіз проблем, що виникають при обробці даних та ефективних інструментів для моделювання та прогнозування даних.
2. Побудовано моделі для аналізу митних надходжень до державного бюджету України за рахунок ввізного та вивізного мита. Було використано мову програмування Python та відповідні до задач пакети.
3. Зазначено можливість використання результатів моделювання фінансово-економічних показників для прийняття управлінських рішень.

Список літератури

[1] У. Маккини, *Python и анализ данных*. М., Россия: ДМК Пресс, 2020, 540 с.
[2] С. Деви, М. Арно, А. Мохамед, *Основы Data Science и BigData. Python и наука о данных*. Петербург, Россия: Питер, 2017, 336 с.
[3] Т. А. Чупілко, "Актуальні проблеми високоефективної обробки даних. Моделювання показників за допомогою мови програмування Python," у *Актуальні напрями розвитку технічного та виробничого потенціалу національної економіки*. Дніпро, Україна: Пороги, 2021, с. 151−163.
[4] Т. А. Чупілко, "Базовий інструментарій у сучасних технологіях комп'ютерної бізнес-аналітики," у *Міжнар. наук.конф. Інноваційні технології, моделі управління кібербезпекою ІТМК-2020*, Дніпро, 2020, т. 2, с. 53−54.
[5] Т. А.Чупілко, "Комп'ютерні технології та економіко-математичні методи в управлінні бізнес-процесами на підприємстві," у *Міжнар. наук. конф. Інноваційні технології, моделі управління кібербезпекою ІТМК-2020*, Дніпро, 2020, Т.1, с. 26−28.
[6] Міністерство фінансів України. [Електронний ресурс]. Режим доступу: http://mof.gov.ua. Дата звернення: 20 серпня, 2021.

Стаття надійшла: 29.08.2021

References

[1] U. Makkyny, *Python y analiz dannykh.* M., Rossia: DMK Press, 2020, 540 p.
[2] S. Devy, M. Arno, A. Mokhamed, *Osnovy Data Science i BigData. Python y nauka o dannkh*. Peterburg, Rossia: Pyter, 2017, 336 p.
[3] T. A. Chupilko, "Aktualni problemy vysokoefektyvnoi obrobky danykh. Modeliuvannia pokaznykiv za dopomohoiu movy prohramuvannia Python," u *Aktualni napriamy rozvytku tekhnichnoho ta vyrobnychoho potentsialu natsionalnoi ekonomiky*. Dnipro: Porohy, 2021, pp. 151−163.
[4] T. A. Chupilko, "Bazovyi instrumentarii u suchasnykh tekhnolohiiakh kompiuternoi biznes-analityky," in *Mizhnar. Nauk. Konf. Innovatsiini tekhnolohii, modeli uprav-linnia kiberbezpekoiu ITMK-2020*, Dnipro, 2020, t. 2, pp. 53−54.
[5] T. A. Chupilko, "Kompiuterni tekhnolohii ta ekonomiko-matematychni metody v upravlinni biznes-protsesamy na pidpryiemstvi," in *Mizhnar. Nauk. Konf. Innovatsiini tekhnolohii, modeli upravlinnia kiberbezpekoiu ITMK-2020*, Dnipro, 2020, t. 1, pp. 26−28.
[6] Ministerstvo finansiv Ukrainu. [Online]. Available: http://mof.gov.ua. Accessed on: August 20, 2021.

Відомості про авторів

Чупілко Тетяна Анатоліївна – кандидат технічних наук, доцент.

Ульяновська Юлія Вікторівна – кандидат технічних наук, доцент, завідувач кафедри.

Мормуль Микола Федорович – кандидат технічних наук, доцент, доцент.

Лагода Анастасія Едуардівна – студентка.

Т. А.Чупилко, Ю. В. Ульяновская, Н. Ф. Мормуль, А. Э. Лагода

PYTHON ДЛЯ ОБРАБОТКИ ДАННЫХ И МОДЕЛИРОВАНИЯ ФИНАНСОВО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Университет таможенного дела и финансов, Днепр

T. A. Chupilko, Yu. V. Ulianovska, M. F. Mormul, A. E. Lagoda

PYTHON FOR DATA PROCESSING AND SIMULATION OF FINANCIAL AND ECONOMIC INDICATORS

University of Customs and Finance, Dnipro

ДО ВІДОМА АВТОРІВ

Найновіші правила оформлення і подання статей знаходяться на сайті журналу
http://itce.vntu.edu.ua/