УДК 343.92 (477):001.18

А.А. Бова,
кандидат соціологічних наук,
старший науковий співробітник

ПРОГНОЗУВАННЯ ЗЛОЧИННОСТІ В УКРАЇНІ: ОДНОВИМІРНІ МОДЕЛІ ЧАСОВИХ РЯДІВ

З метою удосконалення методики прогнозування злочинності розглянуто результати статистичного моделювання динаміки злочинності в Україні протягом 1998–2014 років. Визначено характерні особливості методів прогнозування одновимірних часових рядів. Здійснено прогнозні розрахунки злочинності за допомогою згладжування ковзним середнім, експоненціального згладжування, локально зваженої поліноміальної регресії, авторегресійної моделі, параметри якої ідентифіковані за допомогою методу групового урахування.

Ключові слова: *прогнозування, злочинність, ковзне середнє, експоненціальне згладжування, локально зважена поліноміальна регресія, авторегресійна модель, метод групового урахування аргументів.*

С целью усовершенствования методики прогнозирования преступности рассмотрены результаты статистического моделирования динамики преступности в Украине в течение 1998–2014 годов. Определены характерные особенности методов прогнозирования одномерных временных рядов. Осуществлены прогнозные расчеты преступности с помощью сглаживания скользящим средним, экспоненциального сглаживания, полиномиальной локально-взвешенной регрессии, авторегрессионной модели, параметры которой идентифицированы с помощью метода группового учета.

Ключевые слова: *прогнозирование, преступность, скользящее среднее, экспоненциальное сглаживание, локально-взвешенная полиномиальная регрессия, авторегрессионная модель, метод группового учета аргументов.*

In order to improve crime forecasting techniques the results of statistical modeling of the dynamics of crime in Ukraine during the period of 1998–2014 are reviewed. Features of the methods of forecasting univariate time series are defined. Predictive calculations of crime through moving average smoothing, exponential smoothing, locally weighted polynomial regression, autoregressive model which parameters are identified by using the group method of data handling, are carried out.

Keywords: *forecasting, crime, moving average, exponential smoothing, locally weighted polynomial regression, autoregressive model, group method of data handling.*

Прогнозування злочинності є важливим прикладним завданням кримінології. Кількісні прогнози ґрунтуються на статистичній моделі, що пов'язує історичні зміни часових рядів з майбутнім станом. Під одновимірним часовим рядом ми розуміємо ряд, який складається зі спостережень, зафіксованих через рівні проміжки часу (наприклад, день, тиждень, квартал, рік). Таким чином, часовий ряд характеризується двома змінними – послідовністю набору даних та відповідним часовим періодом. Передбачається єдина методологія спостереження, співставність рівнів рядів динаміки в часі та їхня однозначність за змістом. Існують моделі прогнозу одновимірного часового ряду, які залучають додаткову інформацію про

часовий період (наприклад, підбір форми кривої до емпіричних даних), так і ті, що оперуть лише серією даних (авторегресійний аналіз). Чим точніше статистична модель враховує специфіку складових ряду динаміки (наприклад, наявність тренду та сезонності), тим точнішим буде майбутній прогноз [1]. З кількох статистичних моделей за однакової точності бажано обирати ту, яка містить менше параметрів.

З появою прикладних статистичних та економетричних програм (*STATISTICA*, *XLSTAT*, *GRETL*), а також систем бізнес-аналізу з інтелектуальним вибором методу прогнозування (*Autobox, Forecaster Pro, GMDH Shell*, *KnowledgeMiner Insights*) процес кількісного передбачення став доступним масовому користувачеві. На підставі вивчення сучасної наукової літератури можна виокремити найбільш вживані моделі прогнозування майбутнього стану одновимірного часового ряду:

апроксимація кривої (*Curve Fitting*), наприклад лінійний, логарифмічний, експоненційний, поліноміальний тренд (якщо на графіку візуалізується різка зміна тренду, то можна застосовувати кусочно-лінійні або кусочно-поліноміальні моделі);

декомпозиція часового ряду за трендовою, циклічною та випадковою складовою;

спектральний (Фур'є) аналіз;

локально-зважена поліноміальна регресія (*Locally Weighted Scatterplot Smoothing*, *LOESS* та *LOWESS*);

згладжування ковзним середнім (*Moving Average*) та експоненціальне згладжування (*Exponential Smoothing*);

сингулярний спектральний аналіз ("*Caterpillar*" − *Singular Spectrum Analysis*);

авторегресійна модель ковзної середньої (*Autoregressive Integrated Moving Average*) − методи Бокса−Дженкінса, ОЛІМП.

Крім зазначених статистичних технік, останнім часом до довготривалих спостережень застосовуються адаптивні методи машинного навчання (*Machine Learning*), наприклад, штучні нейронні мережі (*Neural Networks*), дерева рішень (*Decision Trees*), машинна підтримка векторів (*Support Vector Mashine*), генетичне програмування (*Genetic Programming*), багатовимірні адаптивні регресійні сплайни (*Multivariate adaptive regression splines*, *MARS*), регресійний аналіз на попередньо класифікованих даних (алгоритм *M5prim*) тощо. При прогнозуванні коротких рядів динаміки добре зарекомендували себе прогностична модель Грея (*Grey forecasting model*), низка методів математичного моделювання. В умовах невизначеності тенденції можна застосувати метод випадкових блукань (*Random Walk*).

Слід зазначити, що існують й інші статистичні моделі, наприклад факторні, які передбачають включення до регресійного аналізу кілька рядів динаміки, що потребує, зокрема, попереднього прогнозування на наступні роки низки показників демографічного та соціально-економічного розвитку. Але в умовах розгортання соціально-економічної кризи та проведення широкомасштабної антитерористичної операції в Україні кількісні макроекономічні прогнози на віддалений період часу можуть виявитися не дуже достовірними. Кількість облікованих злочинів залежить від багатьох чинників, зокрема, діяльності правоохоронних органів, різких соціальних змін та загострення соціальних суперечностей, поширення в суспільстві певних культурних цінностей тощо. Певну проблему становить достовірність, доступність, повнота та швидкість оновлення статистики. Тому далеко не завжди можливо визначити каузальний зв'язок між деякими чинниками та змінами показників злочинності, знайти адекватну, точну та якісну статистичну модель зв'язку. Одновимірні моделі доцільно завжди застосовувати перед багатовимірним моделюван-

ням часових рядів, аби з'ясувати, наскільки додаткова інформація, наприклад, зміна чисельності населення, його матеріального рівня, соціального розшарування тощо зменшує похибку прогнозу злочинності.

З метою отримання прогнозних значень кількості зареєстрованих злочинів ("КЗ" на рис. 1) було застосовано згладжування ковзним середнім, експоненціальне згладжування, loess – згладжування та модель авторегресії до часового періоду, який охоплював зміну злочинності з 1998 по 2013 рік (постфактум, за даними Генпрокуратури, кількість зареєстрованих злочинів у 2014 р. склала 529139 випадків). Звісно, такий доволі "механічний" аналіз не враховує соціально-економічні процеси, які суттєво впливають на стан злочинності, структурні зміни в останній, трансформацію нормативно-правової бази тощо, а є швидше ілюстрацією деяких можливостей традиційних та сучасних методів прогнозування. Представлений числовий ряд можна трактувати не лише як зміну зареєстрованої злочинності, а й як певні закономірності роботи органів внутрішніх справ з реєстрації та розслідування злочинів. Для ряду характерне протягом кількох років поспіль зростання абсолютного показника зареєстрованої злочинності (1998–1995, 2002–2004) та послідовне зменшення цього показника (1995–2002, 2004–2008).

На підставі критерію Мана–Кендела можна зробити висновок про відсутність тренду в даних, оскільки коефіцієнт кореляції Кандела між рядом зареєстрованої злочинності та періодом спостереження (T=26 років) виявився малим та статистично незначущим (tau=0,01, p=0,95). Ряд динаміки кількості зареєстрованих злочинів (Y_t) описується формулою кубічного поліному (t – номер періоду спостереження).

$$\hat{Y}_t = 100926+212{,}2t^3 - 9567{,}2t^2 + 123256t$$

Коефіцієнт детермінації – квадрат коефіцієнту кореляції між емпіричними значеннями та розрахованими за моделлю – $R^2 = 0{,}87$.

Різниця між емпіричними даними та змодельованими є залишком моделі. Значні індивідуальні значення залишків можуть свідчити про аномальні показники в певний період часу, які не вписуються в модель. На причини, що обумовили такий стан, доцільно додатково звернути увагу. Якість побудованих моделей оцінювалася за допомогою середньої похибки (за модулем), що становить середнє значення модулів різниць між фактичним і передбаченими значеннями часового ряду, розділену на кількість спостережень (*Mean Absolute Error*, *MAE*), та середньої абсолютної похибки прогнозу (за модулем, у %), яка обчислюється на підставі модулю різниці між емпіричним та прогнозованим значенням, тобто абсолютної помилки, поділеної на емпіричне значення в цей момент часу з наступним усередненням отриманих абсолютних процентних помилок (*Mean Absolute Percentage Error*, *MAPE*). Чим менше значення цих показників, тим адекватніше модель відбиває тенденцію зміни емпіричних даних.

Згладжування ковзним середнім. Вилучає випадковості з ряду динаміки (наприклад, помісячна динаміка показника злочинності), прогноз ґрунтується на продовженні даних часового ряду, що згладжується ковзним середнім. Можна взяти більший період усереднення, орієнтуючись на близькість кривих фактичних та прогнозних даних на графіку або числові оцінки різниці між фактичними та згладженими даними. Якщо в даних присутній лінійний тренд, то використовують подвійне ковзке середнє згладжування. Різноманітні види згладжування використовуються зазвичай до серій даних, де періодом спостереження є день чи місяць. Також цю процедуру доцільно проводити перед включенням часових рядів до інших моделей, наприклад, навчання штучних нейронних мереж.

З метою вирівнювання часового ряду та подальшого прогнозування на наступний з трьома періодами усереднення – за два, три та чотири роки. За показниками якості моделі простого однократного ковзного середнього найкращою виявилась модель з періодом згладжування за два роки. На рисунку 1 наведені згладжені дані, що позначені як "КЗ (*MA*)". Прогноз на 2014 р. становить 531598 випадків, що є середнім значенням кількості зареєстрованих злочинів за 2012 та 2013 р. Похибки для зазначеної моделі *MAE*=53534, а *MAPE* = 11. В середньому прогнозні значення відрізняються від реальних на 53534 випадки.

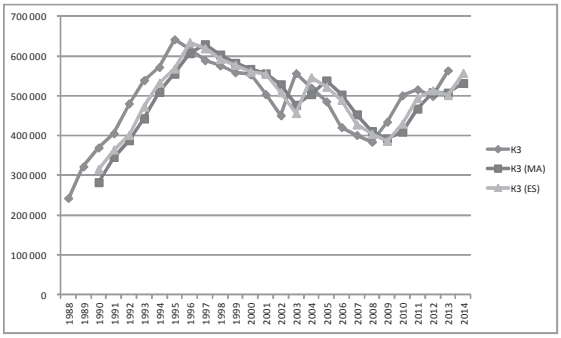

Рис. 1. Початковий та згладжені ряди (ковзним та експоненціальним середнім) кількості зареєстрованих злочинів

Експоненційне згладжування – є аналогом згладжування ковзним середнім, однак при його застосуванні отримуємо зважені за експоненційною функцією значення часового ряду. При експонеціальному згладжуванні враховуються усі попередні спостереження – попередні (останні за часом) з більшими коефіцієнтами, найбільш ранні (початкові) – з меншими коефіцієнтами. Прогнозні значення зручно отримувати на підставі зваженої комбінації фактичного значення з вагою α та його попереднього передбачення з вагою $(1-\alpha)$ за такою рекурентною формулою:

$$S_t=\alpha \times Y_t+(1-\alpha)\times S_{t-1},$$

де S_t – прогнозне значення на часовий період t;

Y_t – значення змінної на часовий період t;

t – часовий період (наприклад, перший, другий, третій місяць року);

α – константа згладжування – фіксований параметр, що знаходиться від 0 до 1;

$(1-\alpha)$ – фактор затухання.

Параметр фактор затухання $(1-\alpha)$ підбирається емпіричним шляхом, орієнтуючись на мінімізацію середнього абсолютного відхилення. Зазвичай його значення становить від 0,2 до 0,3. Значення, з якого ініціюється згладжений ряд (S_0), може бути першим значенням часового ряду або середнім з кількох перших спостережень ряду.

Існує можливість врахування у прогнозі сезонних коливань (двохпараметричне згладжування), тренду (двохпараметричне згладжування за методом Хольта), тренду і сезонних коливань (трьохпараметричне згладжування за методом Вінтерса). За умови більш стрімкої зміни значень часового ряду експонеціальне згладжування дає зазвичай дещо точніші результати прогнозування ніж апроксимація кривої на підставі регресійного аналізу. У ряді випадків прогнозні оцінки, отримані за допомогою експоненціального згладжування, подібні до прогнозів, що ґрунтуються на деяких класах авторегресійних моделей.

Числовий ряд прогнозних значень кількості злочинів на 2014 р. (КЗ "*ES*"), які отримані за допомогою методу експоненціального згладжування (б=0,9), наведено на рисунку 1. Прогноз експоненціального середнього на 2014 р. складається із суми значення за 2013 рік та прогнозованого значення за цей же рік, які взяті з відповідними коефіцієнтами, що становить у підсумку 557307 випадків. Для зазначеної моделі *MAE*=48542, а *MAPE* = 11.

Локально зважена поліноміальна регресія (*Locally weighted polynomial regression*) є різновидом непараметричної регресії, де для кожного загладжуваного значення даних, заданого в точці, обирається набір з фіксованого числа поруч розташованих точок, кожній з яких призначається вага за відповідною формулою. Метод використовується у випадку складної нелінійної залежності, коли вимоги застосування класичного регресійного аналізу не можуть бути виконані, або коли дослідника цікавить насамперед якість прогнозування, а не аналітичний вираз, якому можна дати подальшу змістовну інтерпретацію. Поліноміальна функція стосується лише певного сегменту даних, надаючи більшу вагу найближчим локальним сусідам − спостереженням, які знаходяться поблизу точки, для якої здійснюється оцінка, та меншу вагу − віддаленим точкам. Налаштування моделі включає обрання ступеню поліному (d = 0, d = 1 або d = 2) і параметру згладжування (зазвичай, від q = 0,25 до q = 0,5).

Для заданого локального поліноміального порядку 1 та ширини вікна 0,5 (частка даних, яка використовується для локальної оцінки) на рисунку 2 подано, поряд з оригінальними даними, згладжену криву кількості зареєстрованих злочинів (КЗ (*loess*)). R^2= 0,81, розрахункове значення діянь на 2014 р., за які передбачено кримінальну відповідальність, становить 529019, а похибки прогнозу − *MAE* = 28444, *MAPE* = 10.

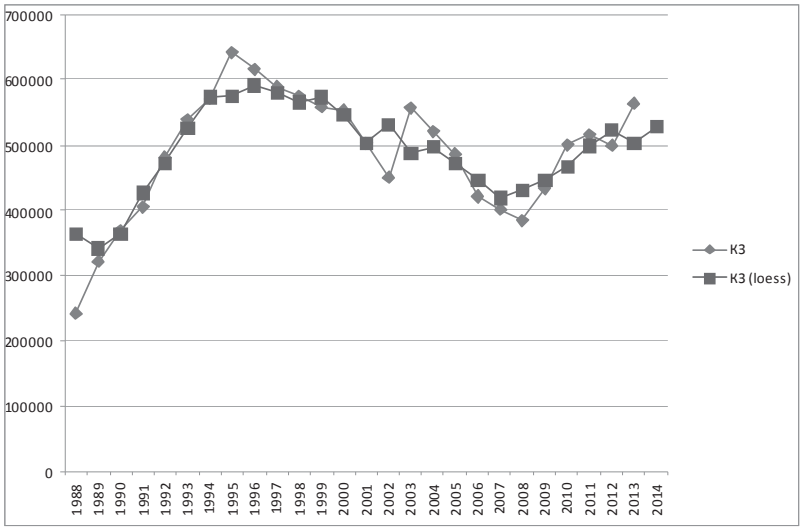

Рис. 2. Початковий та loess-згладжений ряд кількості зареєстрованих злочинів

Авторегресійні моделі визначають залежність значень часового ряду в даний момент від попередніх (не обов'язково сусідніх) значень цього ж ряду. Їх можна використовувати за відсутності або наявності тренду в даних (після вилучення тренду за допомогою різноманітних перетворень часового ряду) та за наявності сезонності. Авторегресійні моделі, попри особливості побудови та оцінку параметрів, мають корисні властивості для аналізу поведінки часового ряду майбутнього прогнозування стану злочинності, оскільки дають можливість

зіставити лаги змінних – показники, що відображають відставання або випередження в часі одного явища порівняно з іншими і на їх основі екстраполювати тенденцію на подальші роки. Звісно, якщо авторегресійна модель ґрунтується на одному–двох лагах, то горизонт прогнозування обмежується одним або двома майбутніми періодами. Порядок лагів авторегресіної моделі можна визначити на підставі автокореляції рівнів ряду. При виборі найкращого рішення з кількох моделей зазвичай обмежуються п'ятьма лаговими змінними. Параметри рівняння можуть буди ідентифіковані різноманітними статистичними методами.

Модель авторегресії – ковзного середнього (*Autoregressive moving average model – ARMA* (*p, q*)) є видом авторегресійної моделі, де для розрахунку модельних значень використовується параметр p – значення змінної в попередні періоди спостереження (*AR*) та q – значення залишків (*MA*). До авторегресійної моделі проінтегрованого ковзного середнього (*Auto-Regressive Integrated Moving Average – ARIMA* (*p, d, q*)) вдаються за умови нестаціонарних рядів динаміки. Якщо в даних присутній тренд, то попередньо обчислюють прирости (d – різниця першого або другого порядку) та проводять розрахунки *ARMA* (*p, q*) з подальшою інтеграцією ряду. Модель може включати додаткові параметри, що задають порядок сезонної складової авторегресії та сезонного ковзного середнього, константу (у випадку її статистичної значущості). Зазвичай для авторегресійних моделей необхідно мати ряд з достатньою кількістю спостережень.

Метод групового урахування аргументів дає можливість включати до аналізу короткі часові ряди, знаходити лінійні та нелінійні рішення авторегресійних процесів, визначати необхідну довжину часового ряду та кількість незалежних змінних для отримання достовірного результату тощо [2]. Деякі алгоритми подають низку релевантних моделей, з яких дослідник може обрати одну, виходячи з попереднього змістовного аналізу (чим ми скористалися нижче), або сформувати колективний прогноз.

При застосуванні зазначеного методу до часового ряду кількості зареєстрованих злочинів (Y_t) ідентифіковані параметри рівняння з п'ятьма незалежними змінними: $X_{1(t-13)}$ – кількість зареєстрованих злочинів, зсунута на тринадцять років назад; $X_{2(t-17)}$ – кількість зареєстрованих злочинів, зсунута на сімнадцять років назад; $X_{3(t-18)}$ – кількість зареєстрованих злочинів, зсунута на вісімнадцять років назад; $X_{4(t-19)}$ – кількість зареєстрованих злочинів, зсунута на дев'ятнадцять років назад; $X_{5(t-21)}$ – кількість зареєстрованих злочинів, зсунута на двадцять один рік назад. Рівняння має такий вигляд:

$$\hat{Y}_t = 1118307{,}5 - 1{,}21X_{1(t-13)} - 0{,}35X_{2(t-17)} + 1{,}6X_{3(t-18)} - 0{,}81X_{4(t-19)} - 0{,}47X_{5(t-21)}$$

Прогноз на 2014 рік за моделлю становить 510009 випадків облікових правопорушень. Помилки прогнозу дорівнюють нулю, а коефіцієнт детермінації (R^2) – одиниці, тобто зв'язок носить функціональний характер. У статистичному сенсі динаміка характеру злочинності протягом останніх років має схожі риси із динамікою 1990-х років.

З проведеного аналізу слід зробити такі висновки.

Прогнозування на підставі одновимірного часового ряду є найбільш економічним з огляду на недостатність додаткової інформації про соціальні процеси, які відбуваються в державі, та їхній взаємний зв'язок. Емпіричні ряди динаміки можуть бути описані різноманітними статистичними моделями. Різні види згладжування та авторегресійних моделей включають значення змінної, яка прогнозується, в попередні роки, що відбиває ефект самодетермінації злочинності.

Попри те, що згладжування ковзним середнім та експоненціальне згладжування застосовується зазвичай до денних або помісячних рядів динаміки, немає жодних пересторог для їхнього використання до річних даних. Представлені моделі вірно вказали на тенденцію зміни зареєстрованої злочинності у 2014 р., однак отримані точкові прогнози моделей різняться між собою та реальним значенням зареєстрованої злочинності.

Наведені в цій роботі та інші методи прогнозування можуть надати додаткову інформацію експерту, який формулює якісні висновки щодо тенденцій розвитку злочинності у країні. Зазначимо, що прогнози можуть бути адекватними за умови незмінності процедури збору даних та сталості тенденції, яка спостерігалася. Минулий рік ознаменувався тимчасовою окупацією частини території, ескалацією конфлікту на сході країни, соціально-економічною кризою. Зрозуміло, що розглянуті методи прогнозів не можуть врахувати різку зміну чисельності населення України, що відбулася між першим та четвертим кварталом 2014 р., наслідки збройного протистояння. Тому в подальших дослідженнях доцільно проводити аналіз за окремими видами злочинів на регіональних даних, з'ясовувати можливі причини розходжень між моделлю та емпіричними даними.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Горчаков А.А.* Математический апарат для инвестора / А.А. Горчаков // Аудит и финансовый анализ. – 1997. – № 3. – С. 1–57.

2. *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем / А.Г. Ивахненко. – К. : Наукова думка, 1981. – 296 с.

Отримано 26.01.2015