РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

2005, том 41, № 1

УДК 519.254

А. Н. Тырсин
(Челябинск)

ИДЕНТИФИКАЦИЯ ЗАВИСИМОСТЕЙ НА ОСНОВЕ МОДЕЛЕЙ АВТОРЕГРЕССИИ

Предложен новый метод идентификации зависимостей на основе моделей авторегрессии. Особенностью метода является не сопоставление выборочных уравнений, а проверка степени близости безразмерных коэффициентов авторегрессии к областям допустимых значений рассматриваемых моделей. Показано соответствие трендовых и авторегрессионных моделей. Приведен пример использования метода для идентификации зависимости расстояния, пройденного автомобилем после подачи сигнала об остановке, от скорости.

Введение. Одной из актуальных задач в технике и экономике при анализе экспериментальной информации является идентификация вида зависимости между зависимой переменной Y (выходной параметр) и одной X или несколькими $\mathbf{X}$ $(X^1, \ldots, X^m)$ независимыми переменными (входные параметры) [1, 2]. Фактически требуется найти функциональную зависимость

$$Y \quad f(\mathbf{X}) \quad , \tag{1}$$

которая наилучшим образом представляет связь между переменными Y и $\mathbf{X}$. Отметим, что – это случайная компонента, характеризующая различие между фактическими значениями y_i зависимой переменной и полученными значениями по модели $\widetilde{y}_i \quad f(\mathbf{X}_i)$.

Уравнение (1) представляет собой регрессионную модель. Часто при анализе процессов бывает трудно выделить основные факторы, влияющие на их развитие. Одним из распространенных подходов к построению зависимостей при этом является представление исследуемого процесса в динамике, т. е. в виде ряда значений исследуемого параметра, полученных в равноотстоящие моменты времени $y_k \quad y(t_k) \quad y(k \quad)$, где – интервал дискретизации. Поэтому здесь вместо переменной x используются моменты времени t. Анализируемый процесс называется временным рядом [3].

Идентификация временного ряда. Для нестационарных рядов (имеющих тренд) традиционный подход к их идентификации основан на методах прямой экстраполяции, использующих различные трендовые модели [4]. Необходимо определить параметры (коэффициенты) одновременно для всех мо-

делей, а затем по одному из критериев выбрать оптимальную. Данный подход имеет недостатки.

1. Для каждой модели анализируется лишь одна выборочная реализация с наилучшими (в некотором смысле) параметрами, что не позволяет формализовать процедуру выбора наилучшей модели. Фактически по параметрическим моделям делается «непараметрический» вывод о моделях в целом.

2. Традиционно используемый F-критерий не позволяет выбрать лучшую из значимых моделей разного функционального вида. Возможны ситуации, когда могут оказаться статистически значимыми несколько различных моделей.

Идентификацию стационарных рядов производят на базе моделей авторегрессии (АР) (или авторегрессии–скользящего среднего) [1], не учитывая физической интерпретации параметров моделей и их структуру. Рассматриваемое решение основано на одновременном использовании экстраполяционных и авторегрессионных моделей, что позволяет построить модель авторегрессии процесса, учитывающую его структуру [5, 6].

В основе метода построения АР-моделей трендовых процессов лежит часто применяемое в теории автоматического управления Z-преобразование временного ряда. Модель авторегрессии имеет вид

$$y_k \quad a_1 y_k \quad _1 \quad a_2 y_k \quad _2 \quad \ldots \quad a_m y_k \quad _m \quad \quad _k, \tag{2}$$

где $a_1, a_2, \ldots, a_m$ – параметры модели АР(m); y_k – значение динамического ряда показателя в k-й момент времени; $_k$ – некий случайный процесс, подаваемый на вход цифрового фильтра с бесконечной импульсной характеристикой порядка m. Наличие ненулевых значений $_k$ характеризует случайную ошибку (рассогласование) между фактическим временным рядом и рядом, построенным по АР-модели (2).

Часто во временном ряду также присутствует аддитивная шумовая компонента, которая приводит к искажению АР-модели. В этом случае вместо модели (2) следует рассматривать модель вида

$$y_k \quad \sum_{i \quad 1}^{m} a_i y_k \quad _i \quad \quad _k;$$

$$u_k \quad y_k \quad \quad _k,$$

где $_k$ – белый шум; u_k – фактически измеренные значения временного ряда.

В работе [7] показано, что наличие аддитивного белого шума приводит к смещению оценок коэффициентов a_i ($i \quad 1, \ldots, m$) относительно своих истинных значений. Одним из возможных способов устранения смещенности является вычитание из элементов главной диагонали информационной матрицы, полученной при использовании метода наименьших квадратов, дисперсии шума.

В данной работе ограничимся случаем, когда аддитивный шум отсутствует. Присутствие $_k$ можно интерпретировать в виде представления оценок коэффициентов a_i ($i \quad 1, \ldots, m$) случайными величинами. Отметим, что такой подход широко распространен при построении регрессионных зависимостей в эконометрике, например, при обосновании замены переменных [2]. В

рамках сделанных допущений определение АР-коэффициентов модели (2) посредством решения задачи минимизации вида

$$a_1, \ldots, a_m \quad \arg\min_{a_i \quad R} \quad y_k \quad a_i y_{k \quad i}^{\quad 2}$$

становится корректным.

Рассмотрим в качестве примера процедуру построения АР-модели линейного тренда

$$y \quad A \quad Bt. \tag{3}$$

Для временного ряда модель линейного тренда (3) примет вид

$$y_k \quad A \quad B(k \quad) \quad A \quad B \quad k, \tag{4}$$

где k – номер отсчета ($k \quad 1,2,\ldots,N$), N – объем выборки (количество отсчетов временного ряда). Применив к последовательности отсчетов (4) Z-преобразование, получим в области изображений

$$Y(Z) \quad \underset{k \quad 0}{} (A \quad B \quad k) z^{\quad k} \quad A \frac{z}{z \quad 1} \quad B \quad \frac{z}{(z \quad 1)^2} \quad \frac{A \quad (B \quad A) z^{\quad 1}}{1 \quad 2z^{\quad 1} \quad z^{\quad 2}}$$

или

$$Y(Z)(1 \quad 2z^{\quad 1} \quad z^{\quad 2}) \quad A \quad (B \quad A) z^{\quad 1}.$$

Выполнив далее для последнего уравнения обратное Z-преобразование, получим в области оригиналов разностную схему

$$y_k \quad 2y_{k \quad 1} \quad y_{k \quad 2} \quad A \quad (k) \quad (B \quad A) \quad (k \quad 1), \tag{5}$$

где (k) – символ Кронекера: $(k) \begin{matrix} 1, & k \quad 0; \\ 0, & k \quad 0. \end{matrix}$ Начиная с $k \quad 2$, $(k) \quad (k \quad 1) \quad 0$, в результате чего формула (5) примет вид $y_k \quad 2y_{k \quad 1} \quad y_{k \quad 2}$.

В работе [8] показано, что между коэффициентами авторегрессии существует вполне определенная взаимосвязь, зависящая от вида модели временного ряда. Это позволяет использовать АР-модели для решения задачи идентификации временных рядов. Идентификацию трендов по значениям АР-коэффициентов можно осуществить на принципах теории распознавания образов. Построение систем распознавания, основанных на реализации данных принципов, использует взаимное пространственное расположение отдельных кластеров. Если кластеры, соответствующие различным классам, разнесены достаточно далеко друг от друга, то применяется такая схема распознавания, как классификация по принципу минимального евклидова расстояния до ближайшего эталона.

Таблица 1

Трендовая модель	Замена переменной	АР-модель	Область допустимых значений АР-коэффициентов
$y_k \quad A \quad Bt_k$	–	$y_k \quad 2y_{k\ 1} \quad y_{k\ 2}$	Точка (2, –1)
$y_k \quad A \quad Bt_k \quad Ct_k^2$	$v_k \quad y_k \quad y_{k\ 1}$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad A \quad \frac{B}{t_k} \quad \frac{C}{t_k^2}$	$v_k \quad y_k t_k^2 \quad y_{k\ 1} t_{k\ 1}^2$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad A \quad \frac{B}{t_k}$	$v_k \quad y_k t_k$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad \frac{1}{A \quad Bt_k}$	$v_k \quad \frac{1}{y_k}$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad \frac{t_k}{A \quad Bt_k}$	$v_k \quad \frac{t_k}{y_k}$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad e^{A \quad B/t_k}$	$v_k \quad t_k \ln y_k$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad Ae^{Bt_k}$	$v_k \quad \ln y_k$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad B \quad Ce^{At_k}$	–	$y_k \quad a_1 y_{k\ 1} \quad a_2 y_{k\ 2}$	$a_1 \quad a_2 \quad 1, \ a_2 \quad 0$
$y_k \quad \frac{A}{1 \quad Ce^{Bt_k}}$	$v_k \quad \frac{1}{y_k}$	$y_k \quad a_1 y_{k\ 1} \quad a_2 y_{k\ 2}$	$a_1 \quad a_2 \quad 1, \ a_2 \quad 0$
$y_k \quad Ct_k e^{Bt_k}$	$v_k \quad \ln \frac{y_k}{t_k}$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)
$y_k \quad \ln(A \quad Bt_k)$	$v_k \quad e^{y_k}$	$v_k \quad 2v_{k\ 1} \quad v_{k\ 2}$	Точка (2, –1)

Рассмотрим табл. 1, которая для наиболее распространенных трендовых моделей определяет область допустимых значений (ОДЗ) в евклидовом векторном пространстве, координатами которого являются АР-коэффициенты.

Замены переменных подбирались таким образом, чтобы минимизировать количество и размеры ОДЗ коэффициентов авторегрессии различных трендовых моделей. Наилучшей моделью признаем ту, для которой расстояние от расчетной точки (вычисленные значения коэффициентов авторегрессии) до соответствующей ОДЗ будет минимальным. Таким образом, минимальное расстояние от параметров модели авторегрессии до множества допустимых значений идентифицирует тренд.

Основная идея метода заключается в том, чтобы множеству всех возможных кривых каждого вида поставить в соответствие некоторую область допустимых значений, не зависящую от значений параметров модели (в частности, одну точку) в евклидовом векторном пространстве, базисом которого являются коэффициенты авторегрессии.

Приведенный в табл. 1 набор моделей, для которых можно построить указанное соответствие, неполный. При необходимости его можно увеличить.

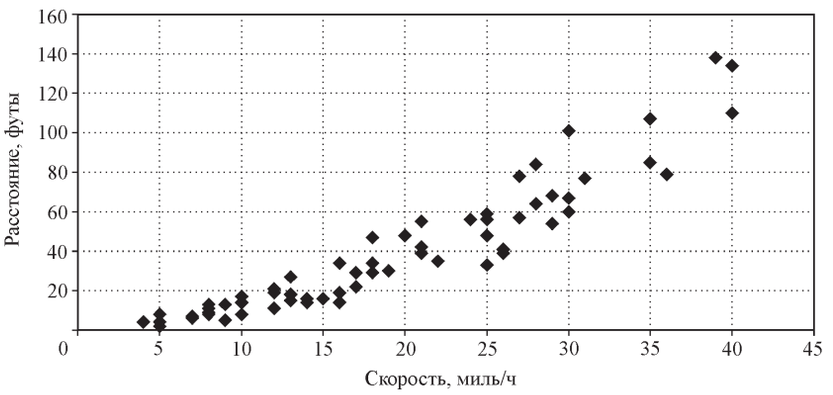

Рис. 1

Идентификация регрессионной зависимости. Идея заключается в равномерной дискретизации пространственной объясняющей переменной, в результате чего получается аналог временного ряда.

Для перехода от зависимости между пространственными переменными к равномерно дискретизированному временному ряду предлагается следующее. Группируем данные с равными интервалами изменений объясняющей переменной. В результате имеем некий аналог временного ряда относительно пространственной объясняющей переменной – пространственный ряд. Вместо времени используется объясняющая переменная. При этом формальная процедура построения модели авторегрессии будет та же самая.

Исходная регрессионная модель заменяется АР-моделью, которая обладает значительно большими возможностями для учета динамики процессов. В частности, это приведет к повышению достоверности прогноза вне анализируемого диапазона независимой переменной для регрессионных моделей. Такой метод рассмотрен в [9, стр. 57]. В предлагаемой работе представлены данные специального эксперимента по изучению зависимости расстояния s, пройденного автомобилем после подачи сигнала об остановке, от его скоро-

Таблица 2

Модель	Расчетное значение F-статистики	Значимость F-статистики
$s \quad b_0 \quad b_1 v$	427,65	$2{,}975 \quad 10^{29}$
$s \quad b_0 \quad b_1 v \quad b_2 v^2$	317,67	$1{,}198 \quad 10^{32}$
$s \quad b_1 v \quad b_2 v^2$	322,19	$8{,}132 \quad 10^{33}$
$s \quad A e^{Bv}$	399,22	$1{,}860 \quad 10^{28}$
$s \quad A v e^{Bv}$	80,19	$1{,}020 \quad 10^{12}$
$s \quad b_0 \quad b_1 \frac{1}{v}$	56,69	$2{,}844 \quad 10^{10}$
$s \quad \frac{1}{b_0 \quad b_1 v}$	49,50	$2{,}012 \quad 10^{9}$

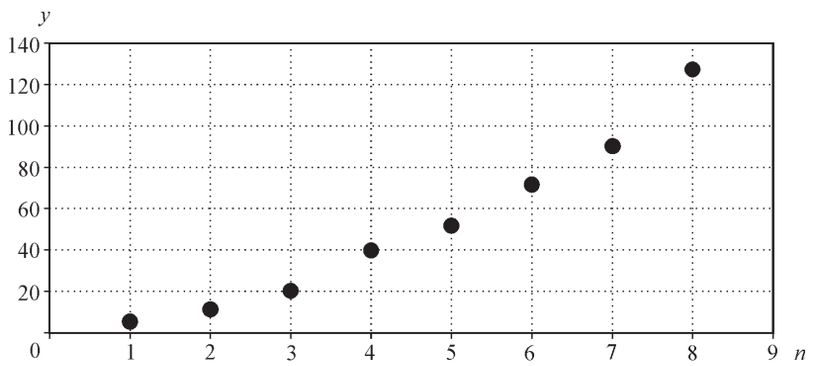

Рис. 2

сти v (рис. 1). В работе [1, стр. 177] показано, что применение известных статистических критериев не позволяет выбрать наилучшую регрессионную модель.

В табл. 2 приведены результаты использования F-критерия Фишера для различных типовых регрессионных зависимостей. Видим, что все модели оказались статистически значимыми. Ввиду разной функциональной формы выбор наилучшей модели оказывается невозможным.

По рассматриваемому методу выполним следующие операции.

1. Сформируем выборку из восьми групп с равными интервалами: (2,5; 7,5), (7,5; 12,5), ..., (37,5; 42,5) миль/ч.

2. Каждому интервалу поставим в соответствие среднегрупповое значение пройденного пути.

3. Сформируем ряд равноотстоящих значений через 5 миль/ч, который соответствует временному ряду, приведенному на рис. 2.

4. Определим АР-коэффициенты для каждой из рассматриваемых моделей.

Результаты расчета приведенных в табл. 1 соотношений даны в табл. 3. Из таблицы видно, что зависимость расстояния, пройденного автомобилем после подачи сигнала об остановке, от его скорости наилучшим образом

Таблица 3

Вид модели	a_1	a_2	Расстояние до ОДЗ
Прямая	1,383	–0,020	1,158
Парабола	0,333	1,170	2,736
Экспонента	1,654	–0,627	0,509
Экспонента + Const	1,383	–0,020	0,257
Экспонента Время	0,866	0,221	1,667
Логистическая	0,836	–0,136	0,213
Обратная	0,836	–0,136	1,449
Гипербола	1,659	–0,119	0,945

представлена логистической моделью. Данная зависимость имеет следующий вид: $s \quad \frac{A}{1 \quad Ce^{Bv}}$, где A, B, C – некоторые постоянные параметры.

Заключение. Рассмотренный метод идентификации регрессионной зависимости является непараметрическим, поскольку реализует выбор наилучшей модели среди заданного множества типовых зависимостей без учета значений их параметров.

АР-модели имеют ряд преимуществ перед традиционными трендовыми и регрессионными моделями: во-первых, количество неизвестных коэффициентов a_i обычно меньше, чем у соответствующих трендовых и регрессионных моделей; во-вторых, АР-модели имеют линейный вид, поэтому их коэффициенты легко вычисляются.

Рассмотренный метод можно использовать при решении задач моделирования, анализе зависимостей и сигналов, а также диагностировании и прогнозировании различных показателей.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. М.: Финансы и статистика, 1985.
2. Клейнер Г. Б., Смоляк С. А. Эконометрические зависимости: принципы и методы построения. М.: Наука, 2003.
3. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, 1974. Вып. 1.
4. Афанасьев В. Н., Юзбашев М. М. Анализ временных рядов и прогнозирование. М.: Финансы и статистика, 2001.
5. Семенычев В. К., Тырсин А. Н. Определение параметров испытательных гармонических сигналов на основе разностных схем // Автометрия. 1991. № 3. С. 95.
6. А. с. 1670464 СССР. Способ определения динамических характеристик линейной механической системы /В. К. Семенычев, А. Н.Тырсин. Опубл. 1991, Бюл. № 30.
7. Прохоров Ю. Н. Статистические модели и рекуррентное предсказание речевых сигналов. М.: Радио и связь, 1984.
8. Тырсин А. Н. Построение трендовых моделей экономических процессов // Сб. науч. ст. междунар. практического семинара-конференции «Проблемы позиционирования российских регионов в мировом экономическом пространстве». Киров, 2002. С. 490.
9. Езекиэл М., Фокс К. Методы анализа корреляций и регрессий. М.: Статистика, 1966.

Челябинский государственный университет,
E-mail: at2001@yandex.ru

Поступила в редакцию
9 февраля 2004 г.