УДК 681.5(075.8)

А.А. Терехов

ИДЕНТИФИКАЦИЯ СТАТИСТИЧЕСКОГО МАТЕРИАЛА И КОНСОЛИДАЦИЯ ВРЕМЕННЫХ РЯДОВ

Излагаются принципы решения задачи идентификации статистического материала с использованием несемантических методов, дано определение алгебраических операций над временными рядами, приведен пример решения задачи консолидации временных рядов и указано ее место в цикле функционирования интеллектуальной системы прогнозирования.

Ключевые слова: *идентификация, консолидация, прогнозирование, формальный процесс, множественный прогноз, тенденция*

Введение

Удовлетворение потребностей органов власти и управления, средств массовой информации, населения, научной общественности, коммерческих организаций и предпринимателей, международных организаций в разнообразной, объективной и полной статистической информации – главная задача Федеральной службы государственной статистики (Росстат). Для решения этой задачи организована система, в состав которой входят центральный аппарат федерального уровня и территориальные органы, расположенные во всех субъектах Российской Федерации.

На официальном сайте Росстата www.gks.ru в разделе «Россия в цифрах» представлено содержание наблюдаемых процессов, которое включает показатели по следующим отраслям: *население, труд, уровень жизни населения, образование, здравоохранение, охрана окружающей среды, правонарушения, валовой внутренний продукт, производство, сельское хозяйство, строительство, транспорт и связь, торговля и услуги населению, финансы, инвестиции, цены, внешнеэкономическая деятельность*. В этом разделе также представлена вся оперативная статистическая информация, которой располагает Росстат. Весь информационный массив регулярно пополняется «свежими» данными и интенсивно используется в административно-хозяйственной деятельности на различных уровнях управления социально-экономическими системами для регулярного решения задач прогнозирования и планирования.

Кроме того, на сайте открыт свободный доступ к центральной базе статистических данных (ЦБСД) и базе данных муниципальной статистики (БДМС). ЦБСД содержит информацию по основным разделам статистики. В состав ЦБСД входит более 2500 показателей годовой, квартальной и месячной периодичности по России, субъектам Российской Федерации, формам собственности, отраслям экономики, видам экономической деятельности и др.

БДМС содержит информацию о статистических показателях муниципальных образований. Состав этих показателей одинаков для всех регионов РФ.

Помимо сбора новых данных, Росстат предоставляет уточняющие сведения за прошедшие периоды как по основным разделам показателей, так и по оперативной информации. При получении уточняющей информации возникает целесообразность повторного прогнозирования актуальных социально-экономических процессов, в результате которого формируются более точные оценки будущих значений показателей, используемых в управленческой деятельности.

Регулярное прогнозирование процессов позволяет не только принимать эффективные управленческие решения, но и накапливать опыт, позволяющий повысить точность и надёжность прогнозов, улучшить методы и алгоритмы прогнозирования.

Многократное решение задач прогноза выявило необходимость создания интеллектуальной системы прогнозирования (ИСП), позволяющей автоматизировать процедуры предсказания, управлять всей совокупностью статистических данных, накапливать опыт и оптимизировать методы прогнозирования конкретных процессов.

В ИСП решаются следующие задачи [2]:

- импорт входных данных – статистического материала;
- идентификация временных рядов, описывающих статистические процессы;
- организация и хранение данных ИСП;
- поиск и навигация по процессам;
- множественное прогнозирование;

- вычисление численных значений характеристик прогнозов;
- консолидация множественных прогнозов;
- адаптация методов прогнозирования и консолидации.

Структурно ИСП состоит из следующих компонентов [2].

Подсистема структуризации и хранения данных, представляющая собой *хранилище,* в котором не только накапливается информация, но при необходимости осуществляется её структурная реорганизация.

Подсистема прогнозирования, которая на основе актуального отрезка временного ряда вычисляет значения множественных прогнозов и их численные характеристики.

Подсистема улучшения алгоритмов прогнозирования и консолидации, которая на основе результатов решения задачи прогноза адаптирует методы прогнозирования и консолидации конкретных процессов.

Подсистема консолидации множественных прогнозов, которая вычисляет конечный (финальный) прогноз на основе множественных прогнозов определенным для данного процесса методом консолидации.

Подсистема управления данными, которая управляет всей совокупностью данных ИСП, осуществляет поиск, навигацию по процессам и их показателям.

Все данные ИСП можно разделить на три вида.

1. Первичные данные – являются входными для ИСП. Это статистический материал, поступающий из различных источников на различных носителях в различных форматах.

2. Вторичные данные – являются внутренними данными ИСП в требуемом формате; они непосредственно используются для получения множественных прогнозов.

3. Третичные данные – формируются в результате решения задач прогноза. Представляют собой значения множественных и финальных прогнозов, их характеристик, сведения об используемых методах прогнозирования и консолидации. На основании третичных данных происходит адаптация и улучшение алгоритмов прогнозирования [2].

Особой задачей подсистемы управления данными является *идентификация* статистического материала. С этой задачей связаны процедуры навигации по процессам, содержащимся в ИСП, и поиска временных рядов, необходимых для решения имеющейся задачи.

Идентификация статистического материала – это выявление наличия в ИСП процесса, представленного временными рядами и метаданными, эквивалентного по некоторым заранее определенным признакам процессу, представленному идентифицируемым статистическим материалом. Если полученный статистический материал описывает уже известный системе процесс, то выделенные из него временные ряды консолидируются операциями дополнения, пересечения и объединения с существующими временными рядами. Если выделенный временной ряд отличается от уже имеющегося в системе шагом наблюдения, то он регистрируется в ИСП как дополнительный временной ряд уже существующего в ИСП процесса.

Если полученный статистический материал не идентифицирован, то в ИСП осуществляется регистрация нового процесса.

Задача идентификации статистического материала возникает по следующим причинам.

1. Регулярность решения задачи прогноза.
2. Регулярность получения статистической информации и необходимость отнесения выделенных из нее временных рядов к одному из имеющихся в системе.
3. Регулярность получения уточненных данных за прошлые периоды.
4. Необходимость поиска и навигации по процессам, показателям и их численным значениям.

Разработка алгоритмов автоматизированной идентификации статистических данных и процедур навигации по процессам позволяет в большей степени автоматизировать решение задачи прогноза и облегчить работу пользователя с ИСП.

Основные понятия и утверждения

Временной ряд является наиболее часто используемым способом представления количественных характеристик статистического процесса. Качественные характеристики процесса представляются с помощью метаданных – лингвистического описания процесса.

Совокупность временного ряда и метаданных, выделенных из исходного статистического материала, представляет *формальный* статистический материал. Формат представления исходного материала может быть разным – электронным, печатным, графическим и т.д. Поэтому в процессе формализации осуществляется приведение входной информации к единому формату, принятому в системе.

Актуальный участок временного ряда, используемый в качестве исходных данных для расчета множественных прогнозов, представляет собой набор данных.

Множественные прогнозы – это множество прогнозов, полученных на основе одного набора данных различными методами. Применяются в многоальтернативном прогнозировании, когда финальный прогноз вычисляется на основании некоторого множества конкурирующих прогнозов.

Финальный прогноз – результирующий, конечный прогноз, вычисленный на основе множественных прогнозов методами консолидации.

Методы консолидации – это методы вычисления значения финального прогноза на основе множественных прогнозов.

Принципы функционирования ИСП

Регулярное решение задачи прогноза обусловливает непрерывное циклическое функционирование ИСП. На вход системы поступает статистический материал, представляющий собой в общем случае разрозненные данные в различных форматах о проблемном процессе. На выходе пользователь получает финальный прогноз.

При поступлении новых или уточненных статистических данных выясняется вопрос: прогнозировались ли ранее эти данные? При положительном ответе на основании сравнения прогноза и факта делаются выводы о приемлемости применяемых методов и моделей прогнозирования проблемного процесса.

Решение задачи прогноза в ИСП состоит из следующих этапов.

1. Получение статистического материала и выделение из него временного ряда и метаданных, представляющих наблюдаемый процессе.
2. Идентификация формального статистического материала.
3. Формирование набора данных – актуального отрезка временного ряда, который будет использоваться при построении модели процесса.
4. Вычисление множественных прогнозов методами прогнозирования.
5. Вычисление значений характеристик множественных прогнозов.
6. Вычисление финального прогноза методами консолидации.
7. Получение нового статистического материала и сопоставление его значений с прогнозными значениями.
8. Вычисление характеристик прогнозов на основе новых статистических данных.

Для каждого этапа в системе ведется журнал событий, где фиксируются все стадии этапа решения задачи прогноза.

Структура формального статистического материала

Исходя из опыта решения задач прогнозирования и работы со статистическим материалом, предоставленным РОССТАТ, можно выделить следующие базовые составляющие метаданных.

1. Источник статистического материала. В нашем случае это РОССТАТ, но помимо этой организации существует большое количество больших (на уровне стран и регионов) и малых (на уровне отраслей, предприятий) статистических организаций.
2. Наблюдаемый процесс/процессы. В большинстве случаев РОССТАТ в рамках одного показателя процесса предоставляет несколько временных рядов с различным шагом измерения (месяц/квартал/год); физически это разные процессы, семантически один и тот же.
3. Условия и критерии наблюдений (РОССТАТ для некоторых процессов в масштабе страны не учитывает данные по Чеченской республики по внутренним соображениям).
4. География наблюдений – географический регион наблюдения процесса.
5. Хронологический период наблюдений – время от начала до конца наблюдений.
6. Шаг наблюдений – хронологическое расстояние между соседними измерениями показателя процесса.
7. Комментарии, пояснения, ссылки и прочее – в предоставляемых документах часто добавляются какие-либо комментарии, пояснения, ссылки на предыдущие документы.

Условные обозначения

Введем условные обозначения.

$P =< \{Y^{[i]}\}, M(P) >$ – проблемный процесс.

$M(P)$ – метаданные, описывающие проблемный процесс.

$Y^{[i]} =< Y_1^{[i]}, Y_2^{[i]}, \ldots, Y_{n_i}^{[i]} >$ – i-й временной ряд,

$Y_j^{[i]}$, $j = \overline{1, n_i}$ – значение i-го временного ряда в момент j.

$\{Y^{[i]}\}$ – множество временных рядов, описывающих процесс P.

$X =< X_1, X_2, \ldots, X_n >$ – новый временной ряд.

$t =< t_0, \Delta t, t_n >$ хронологический интервал наблюдений.

t_0, t_n – соответственно начальная и конечная хронологические даты наблюдений.

Δt – шаг наблюдений.

Хронологические интервалы и шаги наблюдений конкретных процессов указываются при помощи верхних индексов.

$t^{[Y^{[i]}]}$ – хронологический период ряда $Y^{[i]}$.

$t^{[X]}$ – хронологический период ряда X.

$\Delta t^{[Y^{[i]}]}$ – шаг наблюдений временного ряда $Y^{[i]}$.

$\Delta t^{[X]}$ – шаг наблюдений временного ряда X.

Признаки идентификации

В результате анализа различного статистического материала и специфики предметной области были выделены следующие идентификационные признаки:

1) индекс словоформ;
2) источник статистического материала;
3) название описываемого процесса;
4) величина шага наблюдения;
5) хронологический интервал наблюдения;
6) коэффициент совпадения тенденций;
7) коэффициент детерминации;
8) средняя относительная ошибка.

В процессе идентификации используется совокупность данных признаков. Использование каждого признака в отдельности позволяет уменьшить множество выбираемых процессов ИСП. В результате идентификации статистический материал может быть признан идентичным одному из зарегистрированных в ИСП процессов.

Построение индекса словоформ

Обозначим через $I =< W, \vartheta >$ – индекс ключевых словоформ. $W = (W_1, W_2, \ldots, W_n)$ – массив словоформ, $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_n)$ – массив значений частоты повторения словоформ.

Индекс словоформ строится на основе сравнения метаданных, выделенных из статистического материала и метаданных процессов, хранящихся в ИСП. Выделение и анализ словоформ имеют существенное значение для повышения надежности процедуры идентификации.

Неизменяемая часть слова, общая у всех его форм, представляется графической основой, возможно пустой для таких слов как «идти»–«шел». Вся оставшаяся часть слова описывается набором присоединяемых к основе окончаний. Список окончаний, упорядоченных в соответствии с грамматическими формами, образует парадигму (модель) словоизменения.

Определение словоформы реализуется следующим образом. Слово анализируется с конца на совпадение со списком окончаний. Если совпадение имеет место, то окончание отрезается и слово ищется далее в словаре словоформ. Если слово находится – тогда в таблицу соответствий слов описанию заносится номер словоформы слова.

При поступлении новой информации в ИСП производится анализ описания статистических данных. Формируется вектор словоформ, который сравнивается с векторами словоформ процессов, имеющихся ИСП для выявления наиболее близких по данному признаку.

Применение вектора словоформ позволяет существенно уменьшить множество конкурирующих процессов, с которыми сравнивается вновь поступивший статистический материал. Однако он не позволяет сделать однозначный выбор в пользу того или иного процесса.

Источник статистического материала

Источниками статистических данных служат данные федеральных организаций, таких как Федеральное агентство по статистике (РОССТАТ), региональные представительства РОССТАТ, статистические данные аналитических отделов крупных предприятий, ресурсы интернет, СМИ и т.д.

При прочих равных условиях совпадение источников статистических материалов может служить основой для признания их идентичными.

Название описываемого процесса

Различные процессы одинаковой природы могут иметь одинаковые или подобные названия: *численность официально зарегистрированных в службе занятости безработных; численность официально зарегистрированных в службе занятости безработных (на конец периода), численность официально зарегистрированных в службе занятости безработных (на конец периода) в том числе назначено пособие по безработице; общая численность безработных.*

Эти похожие названия соответствуют физически разным процессам с точки зрения статистики. Их временные ряды могут существенно различаться.

Сравнение названия процесса в поставляемом статистическом материале с названиями процессов ИСП увеличивает количество конкурентов, среди которых отыскивается идентичный процесс. За счет этого сокращается вероятность нераспознавания идентичного процесса.

Величина шага наблюдения

Задача идентификации имеет дальнейший смысл решения, если значения шагов наблюдения совпадают. В некоторых случаях при неравных значениях шага задача идентификации имеет смысл, если значения шагов наблюдений являются кратными.

Коэффициент совпадения тенденций

Коэффициент совпадения тенденций показывает степень совпадения тенденций времен-

ных рядов Y и X на соответствующих интервалах отсчета (рисунок 1). Значение коэффициента рассчитывается по следующей формуле

$$p = \frac{h}{l},$$

где h – количество положительных произведений $(Y_{i+1} - Y_i) \cdot (X_{i+1} - X_i)$, $i = \overline{k, k+l}$, l – длина пересекающейся части временных рядов Y и X, т.е. общее количество вычисленных произведений.

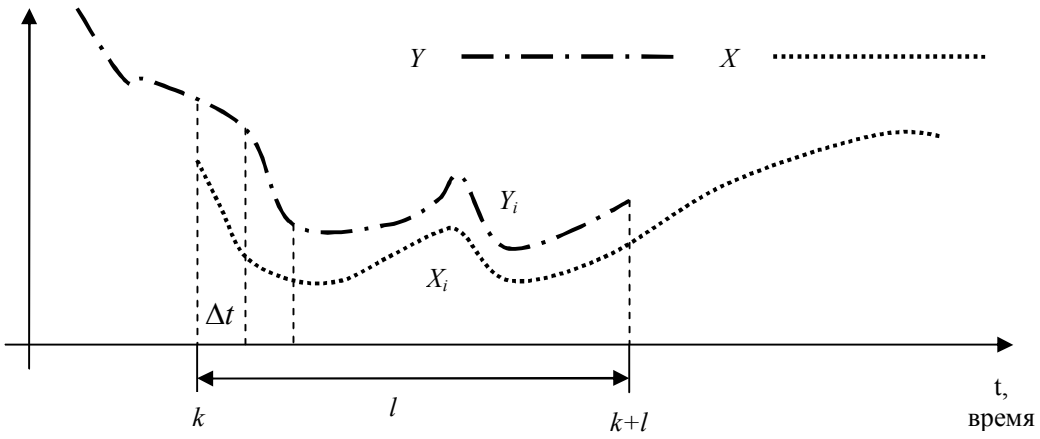

Рисунок 1

Коэффициент детерминации

Коэффициент детерминации – квадрат коэффициента корреляции [3] между сравниваемыми временными рядами Y и X. Он показывает, какая доля вариации Y связана с вариацией X. Коэффициент детерминации не позволяет дать окончательного заключения без учета других признаков, но служит весомым аргументом при принятии решения об идентичности X и Y. Вычисляется по формуле.

$$\eta = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X}) \sum_i (Y_i - \overline{Y})}, \quad (1)$$

где $\overline{Y}$ – среднее значение Y [1].

Среднее относительное отклонение

Относительная ошибка отклонения показывает насколько отличаются соответствующие значения временных рядов Y и X. Вычисляется по формуле

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{\sum_{i=1}^{n} (Y_i)^2}}, \quad (2)$$

Использование рассмотренные признаков идентификации временных рядов не требует применения методов искусственного интеллекта.

Определение алгебраических операций над временными рядами

Y, X – временные ряды. $Y_i, i = \overline{1, n}$ – порядковые номера значений временного ряда Y; $X_j, j = \overline{1, m}$ – порядковые номера значений временного ряда X. Наложим ограничение, что Y и X эквидистантные временные ряды [4], т.е. значения временных промежутков между соседними значениями временного ряда равны: $\Delta t^{[X]} = \Delta t^{[Y]}$. Обозначим их Δt. Помимо порядковых номеров значений для каждого временного ряда, которые вводятся для упрощения работы, существуют хронологические. Они обозначены как $t^{[Y]}$ и $t^{[X]}$. $t_0^{[Y]}$ и $t_n^{[Y]}$ – соответственно начальное и конечное значения хронологического интервала ряда Y. $t_0^{[X]}$ и $t_n^{[X]}$ – соответственно начальное и конечное значения хронологического интервала ряда X. Для удобства определения алгебраических операций введем следующее неравенство

$$t_0^{[Y]} < t_0^{[X]} < t_n^{[Y]} < t_n^{[X]}.$$

Расположив временные ряды хронологически на временной оси друг относительно друга (рисунок2), дадим определение алгебраических операций над временными рядами.

Введем операцию нечеткого равенства $\cong$ временных рядов. Временные ряды X и Y нечетко равны $Y \cong X$, если их хронологические интервалы совпадают, и выполняется неравенство:

$$\frac{\sum_{i=1}^{n} |X_i - Y_i|}{n} < \varepsilon, \quad (3)$$

где ε – допустимое значение ошибки, задаваемое априорно на основе экспертной оценки.

Операция нечеткого равенства вводится с целью ослабления жесткого равенства соответствующих значений сравниваемых временных рядов. Любая статистическая информация имеет погрешность с реальными, фактическими значениями. Какими бы точными не были бы методы сбора физической информации, методы расчета значений, погрешность всегда имеет место. Поэтому точное равенство заменим нечетким равенством.

Объединением временных рядов Y и X – обозначается $Y \cup X$ – является такой ряд Z, $Z_k, k = \overline{1, L(t_Y \cap t_X)}$, который определяется следующим образом (рисунок 2). $L(t^{[Y]} \cap t^{[X]})$ – длина временного интервала, образованного пересечением временных интервалов $t^{[Y]} \cap t^{[X]}$.

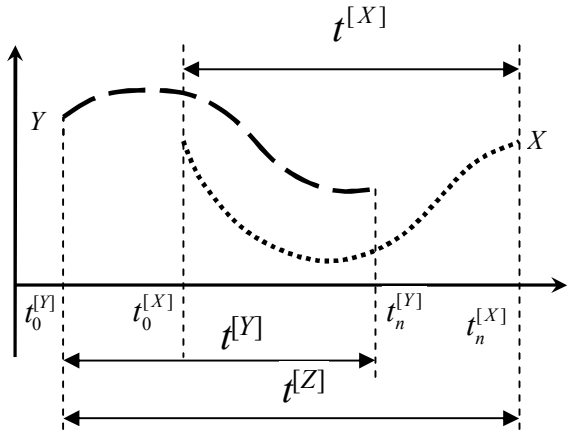

Рисунок 2

Для операции конкатенации будет использован символ $\&$. Временной интервал $t^{[Z]}$ будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]}; t_0^{[X]}]$, $(t_0^{[X]}; t_n^{[Y]})$, $[t_n^{[Y]}; t_n^{[X]}]$. Значения временного ряда Z на отрезке $[t_0^{[Y]}; t_0^{[X]}]$ будут равны соответствующим значениям временного ряда Y на интервале $[t_0^{[Y]}; t_0^{[X]}]$ – $Z_{[t_0^{[Y]}; t_0^{[X]}]} = Y_{[t_0^{[Y]}; t_0^{[X]}]}$. Значение временного ряда Z на интервале $(t_0^{[X]}; t_n^{[Y]})$ – $Z_{(t_0^{[X]}; t_n^{[Y]})}$, для удобства его можно обозначить отрезком $[t_0^{[X]} + \Delta t; t_n^{[Y]} - \Delta t]$ в виду дискретности временного ряда, т.е. $Z_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]}$, вычисляется как среднее значение между X и Y. $Z_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]} = \overline{Y_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]}, X_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]}}$. Значения временного ряда Z на отрезке $[t_n^{[Y]}; t_n^{[X]}]$ будут равны соответствующим значениям временного ряда X на интервале $[t_n^{[Y]}; t_n^{[X]}]$ – $Z_{[t_n^{[Y]}; t_n^{[X]}]} = X_{[t_n^{[Y]}; t_n^{[X]}]}$.

$$Z_{[t_n^{[Y]}; t_n^{[X]}]} = Y_{[t_0^{[Y]}; t_0^{[X]}]} \,\&\, \overline{Y_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]}, X_{[t_0^{[X]}+\Delta t; t_n^{[Y]}-\Delta t]}} \,\&\, X_{[t_n^{[Y]}; t_n^{[X]}]}, \quad (4)$$

где черта сверху означает вычисление среднего арифметического соответствующих элементов усредняемых рядов.

Пересечением временных рядов Y и X – обозначается $Y \cap X$ – является ряд Z, $Z_k, k = \overline{1, l}$, $l = L([t_0^{[X]} : t_n^{[Y]}])$ – длина пересечения временных рядов Y и X (рисунок 3).

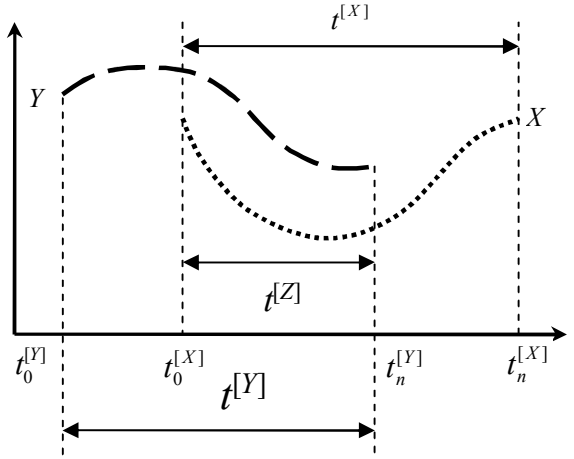

Рисунок 3

Временной ряд Z определен только на отрезке $[t_0^{[X]}; t_n^{[Y]}]$. Значения ряда вычисляются как среднее между соответствующими значениями X и Y.

$$Z_{[t_0^{[X]}; t_n^{[Y]}]} = \overline{Y_{[t_0^{[X]}; t_n^{[Y]}]}, X_{[t_0^{[X]}; t_n^{[Y]}]}}, \quad (5)$$

Дополнение слева временного ряда X рядом Y – обозначается $X_{\to X} \to Y$ – является такой ряд Z, который определяется следующим образом (рисунок 2).

Временной интервал t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]}; t_0^{[X]})$, $[t_0^{[Y]}; t_n^{[X]}]$. Значения временного ряда Z на интервале $[t_0^{[Y]}; t_0^{[X]})$, для удобства обозначим его отрезком $[t_0^{[Y]}; t_0^{[X]} - \Delta t]$ в виду дискретности временного ряда, т.е. $Z_{[t_0^{[Y]}; t_0^{[X]}-\Delta t]}$ будут равны соответствующим значениям временного ряда Y на отрезке $[t_0^{[Y]}; t_0^{[X]} - \Delta t]$ – $Z_{[t_0^{[Y]}; t_0^{[X]}-\Delta t]} = Y_{[t_0^{[Y]}; t_0^{[X]}-\Delta t]}$. Значения временного ряда Z на интервале $[t_0^{[X]}; t_n^{[X]}]$ – $Z_{[t_0^{[X]}; t_n^{[X]}]}$, будут равны соответствующим значениям временного ряда X на интервале $[t_0^{[X]}; t_n^{[X]}]$ –

$Z_{[t_0^{[X]};t_n^{[X]}]} = Y_{[t_0^{[X]};t_n^{[X]}]}$.

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_0^{[X]}-\Delta t]} \,\&\, X_{[t_0^{[X]};t_n^{[X]}]}, \quad (6)$$

Дополнением справа временного ряда Y рядом X – обозначается $Y_{Y\leftarrow} \to X$ – является такой ряд Z, который определяется следующим образом (рисунок 3).

Временной интервал t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]};t_n^{[Y]}]$, $(t_n^{[Y]};t_n^{[X]}]$. Значения временного ряда Z на интервале $[t_0^{[Y]};t_n^{[Y]}]$ $Z_{[t_0^{[Y]};t_n^{[Y]}]}$ будут равны соответствующим значениям временного ряда Y на отрезке $[t_0^{[Y]};t_n^{[Y]}]$ – $Z_{[t_0^{[Y]};t_n^{[Y]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]}$. Значения временного ряда Z на интервале $(t_n^{[Y]};t_n^{[X]}]$, для удобства обозначим его отрезком $[t_n^{[Y]}+\Delta t;t_n^{[X]}]$ в виду дискретности временного ряда, т.е. $Z_{[t_n^{[Y]}+\Delta t;t_n^{[X]}]}$ будут равны соответствующим значениям временного ряда X на отрезке $[t_n^{[Y]}+\Delta t;t_n^{[X]}]$ – $Z_{[t_n^{[Y]}+\Delta t;t_n^{[X]}]} = Y_{[t_n^{[Y]}+\Delta t;t_n^{[X]}]}$

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]} \,\&\, X_{[t_n^{[Y]}+\Delta t;t_n^{[Y]}]}, \quad (7)$$

$t_0^{[Y]} < t_n^{[Y]} \le t_0^{[X]} < t_n^{[X]}$

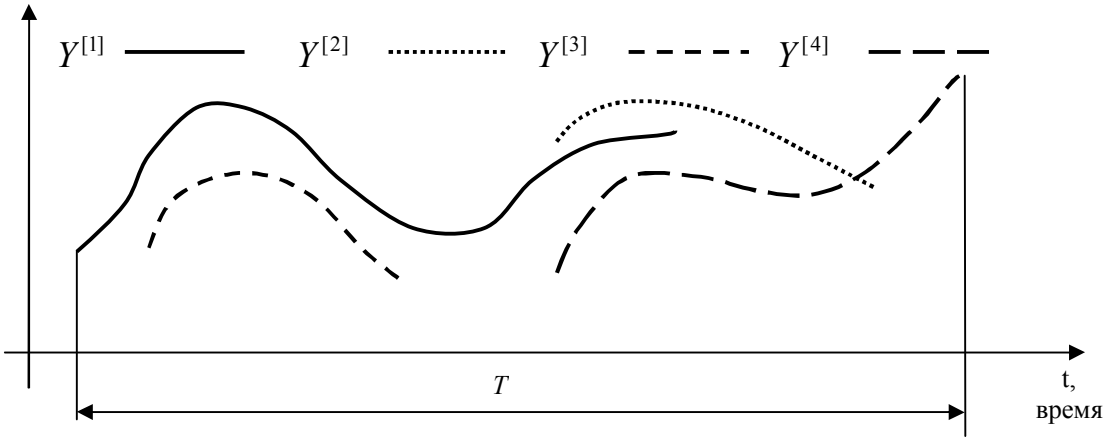

Рисунок 4

Ситуации, возникающие в процессе решения задачи идентификации статистического материала

В общем случае статистический процесс описывается совокупностью временных рядов. Эти ряды могут быть получены из различных источников, иметь различные хронологические интервалы, разные шаги наблюдений и т.д., но семантически будут относиться к одному процессу.

На рисунке 4 изображен общий случай описания процесса $P =< \{Y^{[i]}\}, M(P) >$ временными рядами $Y^{[1]}$, $Y^{[2]}, Y^{[3]}, Y^{[4]}$. В случае положительного решения задачи идентификации – т.е. временные ряды $Y^{[1]-[4]}$ относятся к процессу $P =< \{Y^{[i]}\}, M(P) >$ – возникает задача определения единого временного ряда на временном интервале T (рисунок 4). Общий случай при этом разбивается на частные с использованием алгебраических операций над временными рядами, описанными выше. Рассмотрим следующие варианты частных случаев.

Вариант 1. Представлен на рисунке 3 – $t_0^{[Y]} < t_0^{[X]} < t_n^{[Y]} < t_n^{[X]}$. Временной ряд X относительно ряда Y на временной оси хронологически расположен правее или позднее, т.е. в случае идентичности хранит в себе более «свежие» данные. Если эти ряды описывают один и тот же процесс, о чем может говорить прежде всего, нечеткое равенство $Y_{[t_0^{[X]};t_n^{[Y]}]} \cong X_{[t_0^{[X]};t_n^{[Y]}]}$, то для дальнейшего функционирования системы прогнозирования и обновления данных о проблемном процессе выполнить операцию *дополнения справа Y рядом X*, *дополнения слева ряда X* рядом Y или *объединения* временных рядов X и Y. Вид операции определяется в зависимости от степени точности, актуальности и новизны временных рядов X и Y.

Вариант 2. $t_0^{[Y]} < t_n^{[Y]} < t_0^{[X]} < t_n^{[X]}$ (рисунок 5). Если по признакам идентификации 1-4 ряды X и Y описывают один и тот же процесс, то в зависимости от соотношения значений d и Δt величины возможны следующие случаи.

- $d < 2 \cdot \Delta t$ – допущена погрешность при расчете значений временного ряда

(запаздывание полученной информации, ошибка в дате расчет и т.д.). Временной интервал ряда t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]};t_n^{[Y]}]$ и $[t_0^{[X]};t_n^{[X]}]$

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]} \,\&\, X_{[t_0^{[X]};t_n^{[X]}]}, \qquad (8)$$

- $d \approx 2 \cdot \Delta t$ – пропущено одно значение временного ряда, пропущенный «аргумент-дата» в этом случае равен t^*. Временной интервал t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]};t_n^{[Y]}]$, $[t_n^{[Y]};t_0^{[X]}]$, $[t_0^{[X]};t_n^{[X]}]$. Обозначим временной ряд на участке $[t_n^{[Y]};t_0^{[X]}]$ через R, который имеет три значения $R_{t_n^{[Y]}}$, $R_{t^{[*]}}, R_{t_0^{[X]}}$. Значения $R_{t_n^{[Y]}} = Y_{t_n^{[Y]}}$, $R_{t_0^{[X]}} = Y_{t_0^{[X]}}$. Значение $R_{t^{[*]}}$ может быть получено от источника статистического материала. Если на данный момент это невозможно, то для целей решения задачи прогноза данное значение восстанавливается путем экстраполяции

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]} * R_{[t_n^{[Y]};t_0^{[X]}]} * X_{[t_0^{[X]};t_n^{[X]}]}, \qquad (9)$$

- $d > 2 \cdot \Delta t$ – пропущено более одного значения. Аналогично предыдущему пункту временной интервал t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]};t_n^{[Y]}]$, $[t_n^{[Y]};t_0^{[X]}]$, $[t_0^{[X]};t_n^{[X]}]$. Временной ряд R будет состоять более чем из трех значений. Если значения данного процесса за временной период $[t_n^{[Y]};t_0^{[X]}]$ не могут быть получены от какого-либо источника, то при наличии возможности (малая длина R относительно длины рядов Y и X) они восстанавливаются путем экстраполяции,

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]} * R_{[t_n^{[Y]},t_0^{[X]}]} * X_{[t_0^{[X]};t_n^{[X]}]}, \qquad (10)$$

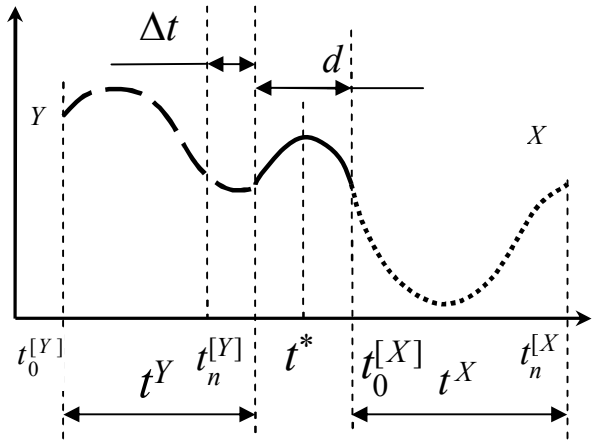

Рисунок 5

Если по признакам идентификации 1-4 ряды Y и X описывают разные процессы, то в системе заводится новый процесс, описываемый временным рядом X.

Вариант 3. $t_0^{[Y]} < t_n^{[Y]} \approx t_0^{[X]} < t_n^{[X]}$ (рисунок 6. Это частный случай описанного в *варианте 2*, с той разницей, что $d \approx 0$. Если по признакам идентификации 1-4 ряды Y и X описывают один и тот же процесс, то временной интервал t_Z будет представлять собой конкатенацию временных интервалов $[t_0^{[Y]};t_n^{[Y]}]$ и $[t_0^{[X]};t_n^{[X]}]$, причем $t_m^{[Y]} \approx t_0^{[X]}$

$$Z_{[t_0^{[Y]};t_n^{[X]}]} = Y_{[t_0^{[Y]};t_n^{[Y]}]} * X_{[t_0^{[X]};t_n^{[X]}]}, \qquad (11)$$

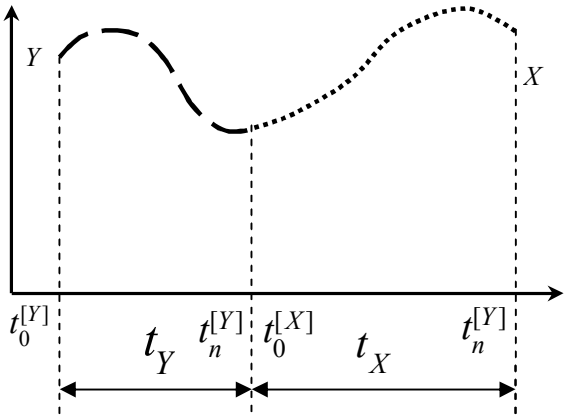

Рисунок 6

Вариант 4. $t_0^{[Y]} \approx t_0^{[X]} < t_n^{[Y]} \approx t_n^{[X]}$ (рисунок 7). Возможны следующие случаи.

- Если по признакам идентификации 1-6 ряды X и Y описывают один и тот же процесс, необходимо акцентировать внимание на признаке семь: средняя относительная ошибка. Если ее значение не превышает некоторый порог априорной погрешности представления данных, делаем вывод, что эти временные ряды описывают один и тот же процесс. Для дальнейшего решения задачи прогноза выполняется операция пересечения временных рядов или оба временных ряда заносятся в систему, прогноз выполняется на основании двух рядов и затем консолидируется.
- Если по признакам идентификации 1-6 ряды X и Y описывают один и тот же процесс, но значение средней относительной ошибки превышает некоторый порог априорной погрешности, делаем вывод, что данные временные ряды описывают семантически разные процессы. В системе определяется новый процесс и временной ряд X ассоциируется с ним.
- Если по признакам идентификации 1-6 ряды X и Y описывают различные процессы, то в системе определяется новый про-

цесс и временной ряд X ассоциируется с ним.

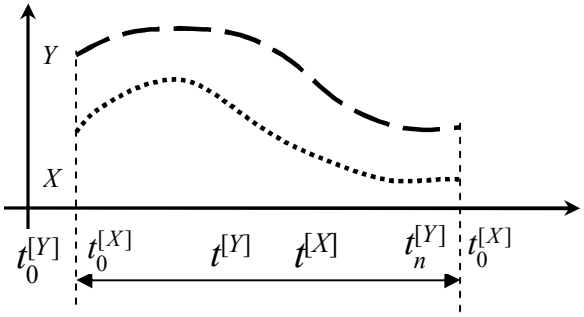

Рисунок 7

Формирование консолидированного временного ряда

Консолидация временных рядов – вычисление единого обобщенного временного ряда на основе нескольких различных.

В процессе решения задачи консолидации временных рядов $Y^{[1]}$, $Y^{[2]}$, $Y^{[3]}$, $Y^{[4]}$ временной интервал T делится на три интервала T_1, T_2, T_3 (рисунок 8). На каждом интервале над временными рядами выполняется одна из алгебраических операций, описанных выше. В результате выполнения этих операций в случае, приведенном на рисунке 4, получаем три временных ряда, расположенных в хронологической последовательности аналогично *Варианту 3*.

Результатом выполнения операции объединения временных рядов $Y^{[1]}, Y^{[2]}, Y^{[3]}, Y^{[4]}$ является временной ряд $Y^{[*]}$, представленный на рисунке 8.

Вычисление консолидированного временного ряда может выполняться в том случае, когда процессы и их временные ряды признаны идентичными. Если вновь поступившему процессу не найден идентичный в ИСП, то этот процесс регистрируется в системе как новый и задача консолидации временных рядов не выполняется.

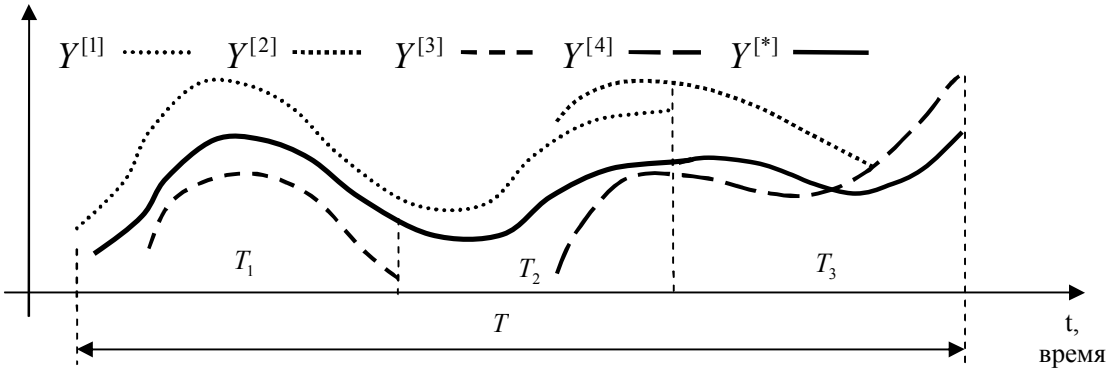

Рисунок 8

Заключение

Основные теоретические результаты

Введены понятия: формальный статистический материал; формальный процесс, консолидированный временной ряд, консолидированный и множественный прогнозы.

Определены алгебраические операции над временными рядами, позволяющие согласовать новый и существующие временные ряды для процессов, идентификация которых прошла успешно.

Определены качественные и количественные идентификационные признаки статистического материала, включая индекс словоформ и коэффициент совпадения тенденций.

Определены принципы функционирования интеллектуальной системы прогнозирования.

Прикладное значение полученных результатов

Изложенные в настоящей статье теоретические положения позволяют формализовать процедуры идентификации статистического материала и консолидации временных рядов, что позволяет минимизировать количество потенциальных ошибок в процессе реализации системы регулярного прогнозирования.

Библиографический список

1. *Айвазян С.А., МхитарянВ.С.* Прикладная статистика и основы эконометрики. – М.:ЮНИТИ, 1998. – 1024 с.

2. *Терехов А.А*. Принципы идентификации временных рядов системы регулярного прогнозирования структурно-неустойчивых процессов // Новые информационные технологии в научных исследованиях и в образовании НИТ – 2007 12 Всероссийская НТК студентов, молодых учёных и специалистов. Тезисы докладов. – Рязань: РГРТА. 2007. – С. 9 – 11.

3. *Афанасьев В.Н.*, Юзбашев М.М. Анализ временных рядов: учебник – М.: Финансы и статистика, 2001. – 228 с.

4. *Тюрин Ю.Н.*, Макаров А.А. Анализ данных на компьютере – Под ред. В.Э. Фигурнова. – 3-е изд., перераб. и доп. – М.: ИНФРА-М, 2003. – 544 с., ил.