ПОСТРОЕНИЕ ЭКСПЕРТНО-СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ПО НЕПОЛНЫМ ДАННЫМ

DOI: 10.36724/2072-8735-2021-15-6-33-39

Manuscript received 20 January 2021;
Revised 18 February 2021;
Accepted 22 March 2021

Носков Сергей Иванович,
Иркутский государственный университет путей сообщения, г. Иркутск, Россия,
sergey.noskov.57@mail.ru

Ключевые слова: *регрессионная модель, неопределенность в данных, пропуски, интервальная система линейных алгебраических уравнений, квазирешения, оценки параметров*

Рассматривается проблема построения линейной регрессионной модели по неполным данным, содержащим пропуски, с привлечением статистической и экспертной информации. Причинами пропусков в данных могут быть, в частности, временная неисправность (сбой) измерительной аппаратуры при снятии различных технических характеристик, или небрежность в работе статистических служб при фиксации отчетных показателей. Весьма часто пропуски возникают при обработке разного рода социологической информации, имеющей форму анкет, когда респонденты отказываются отвечать на какой-то конкретный вопрос (но отвечают на другие) или дают недопустимый, в частности, уклончивый ответ. Предлагаемый в работе подход предполагает заполнение пропусков интервалами, границы которых формируют эксперты, руководствуясь при этом как своими опытом и знаниями об объекте исследования, так и привлекая известные методы точечного заполнения пропусков. После этого оценивание параметров модели в зависимости от характера исходной неопределенности в данных сводится к решению задач линейного или частично-булевого линейного программирования. Рассмотрен случай, когда решение формализующей неопределенность в исходных данных интервальной системы линейных алгебраических уравнений не является единственным. Решена задача построения линейного регрессионного уравнения влияния объема экспорта крупнотоннажных контейнеров и грузооборота железнодорожного транспорта КНР на объем импорта крупнотоннажных контейнеров на железнодорожном пункте пропуска Забайкальск-Маньчжурия.

Информация об авторе:
Носков Сергей Иванович, *д.т.н., профессор кафедры "Информационные системы и защита информации", Иркутский государственный университет путей сообщения, г. Иркутск, Россия*

Для цитирования:
Носков С.И. Построение экспертно-статистических моделей по неполным данным // T-Comm: Телекоммуникации и транспорт. 2021. Том 15. №6. С. 33-39.

For citation:
Noskov S.I. (2021) Construction of expert-statistical models from incomplete data. *T-Comm*, vol. 15, no.6, pp. 33-39. (*in Russian*)

Введение

Один из наиболее эффективных подходов к моделированию сложных объектов различной природы с целью их анализа и прогнозирования основывается на методах современной прикладной статистики, совокупность которых принято называть также «анализом данных». Традиционная схема применения этого подхода предполагает формализованное описание функционирования объекта в прошлом с привлечением для обработки статистической или экспериментальной информации соответствующих специальных методов экстраполяционного характера на этапах формирования модельных спецификаций и оценивания неизвестных параметров. Процесс прогнозирования при этом состоит в варьировании значений внешних переменных на периоде упреждения прогноза в рамках заданных сценариев с расчетом соответствующих значений внутренних переменных модели.

Как отмечено в [1], такой подход вполне обоснован и оправдан при исследовании методами математического моделирования хорошо изученных объектов, функционирование которых подчиненно устойчивым закономерностям на периоде основания прогноза, а степень инерционности процесса такова, что нет оснований предполагать их нарушения на периоде упреждения прогноза и, кроме того, если вся исходная информация, включая ретроспективную и прогнозную (об экзогенных переменных), полностью определена.

Вместе с тем, при исследовании многих сложных систем, в частности, социально-экономических, эти предпосылки нарушаются, а именно:

– степень изученности объекта не позволяет при описании его функционирования ограничиться только формальными средствами, не привлекая знаний специалистов в данной предметной области на различных этапах исследования;

– присущие объекту внутренние закономерности функционирования претерпевают значительные изменения уже в ретроспективном периоде;

– тенденции функционирования объекта на периодах основания и упреждения прогноза не совпадают;

– имеет место неопределенность в статистической, прогнозной и (или) экспертной информации.

Применение в рамках анализа данных только «классической» методологии моделирования сложных объектов, характеризующихся указанными свойствами, оказывается явно недостаточным для построения математических моделей, вполне адекватных исследуемым процессам. Необходима значительная адаптация известных и разработка новых методов для создания качественных моделей таких объектов. При этом если модель, построенная только на основе полной ретроспективной информации об объекте, принято называть «статистической», то модель, разработанную с привлечением наряду с методами регрессионного анализа еще и экспертной информации, естественно считать «экспертно-статистической моделью» (ЭСМ) (см., например, [2]).

Методам построения ЭСМ посвящено значительное количество работ. Так, в [3] предложен способ выбора класса линейной по параметрам регрессии на основе экспертных высказываний. В [4] рассмотрены способы комбинирования (сочетания) прогнозов с учетом экспертной информации относительно возможного изменения тенденций функционирования исследуемого объекта в будущем по отношению к периоду основания. В работе [5] предложен метод оценивания параметров математической модели регрессионного типа одновременно по статистической и экспертной информации, основанный на использовании метода наименьших модулей (МНМ) и сводящийся к задачам линейного программирования (ЛП). Наконец, в работе [6] решена задача прогнозирования эндогенной переменной на основе дискретной динамической модели с использованием специальным образом сформированной экспертной информации. В [7] предложен метод аналогов в прогнозировании коротких временных рядов при построении ЭСМ. Работа [8] посвящена построению регрессионных моделей для описания данных, разные части которых получены из разных источников и представлены в разных шкалах. Такие данные могут возникать, например, при комбинировании фактографической и экспертной информации. При этом изучаются случаи, когда непрерывные данные дополняются порядковыми и номинальными данными. Модели образуются комбинированием (с использованием совместной параметризации) известных моделей для соответствующих типов шкал.

Настоящая работа посвящена решению проблемы построения регрессионных моделей в случае неполноты исходной информации и ее восполнения посредством привлечения информации экспертного характера.

1. Постановка задачи

Рассмотрим линейное регрессионное уравнение

$$y_k = \sum_{i=1}^{m} a_i x_{ki} + \varepsilon_k , \qquad k = \overline{1,n} \tag{1}$$

где y – зависимая, x_i – i-ая независимая переменные; a_i – i-ый оцениваемый параметр; ε_k – ошибки аппроксимации, k – номер наблюдения, n – число наблюдений (длина выборки).

Представим уравнение (1) в векторной форме:

$$y = Xa + \varepsilon \tag{2}$$

где $y = (y_1, ..., y_n)^T$, $a = (a_1, ..., a_m)^T$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$, $X - (n \times m)$ – матрица с компонентами x_{ki}.

В традиционной постановке [9] все элементы матрицы X и вектора y считаются заданными. Однако возможны ситуации, когда это требование нарушается, и выборка данных – пара (X,y), – содержит пропуски. Формально это может быть выражено следующим образом.

Введем в рассмотрение матрицу $A^X = \left\| a_{ki}^x \right\|$, $k = \overline{1,n}$, $i = \overline{1,m}$ и вектор $b^y = (b_1^y, b_2^y, ..., b_n^y)$ индикаторов пропусков следующим образом:

$$a_{ki}^x = \begin{cases} 1, \text{если элемент } x_{ki} \text{ задан числом} \\ 0, \text{если элемент } x_{ki} \text{ не определен,} \end{cases}$$

$$b_{ki}^x = \begin{cases} 1, \text{если элемент } y_k \text{ задан числом} \\ 0, \text{если элемент } y_k \text{ не определен.} \end{cases}$$

При этом величина $\theta = (\sum_{k=1}^{n}(\sum_{i=1}^{m} a_{ki}^{x} + b_{k}^{y}))/(mn+n)$ может служить мерой определенности данных.

Причинами пропусков в данных могут быть, в частности, временная неисправность (сбой) измерительной аппаратуры при снятии различных технических характеристик, или небрежность в работе статистических служб при фиксации отчетных показателей. Весьма часто пропуски возникают при обработке разного рода социологической информации, имеющей форму анкет, когда респонденты отказываются отвечать на какой-то конкретный вопрос (но отвечают на другие) или дают недопустимый, в частности, уклончивый ответ.

Некоторые исследователи предлагают исключать из рассмотрения наблюдения с пропусками и оставшуюся, "комплектную", выборку обрабатывать традиционным образом. Однако при анализе так называемых малых или средних выборок такое искусственное их «усечение» является слишком расточительным с точки зрения потери полезной информации, как совершенно справедливо отмечается в [10].

К настоящему времени разработано большое число методов заполнения пропусков данных. Наиболее известными из них следует, видимо, считать давно разработанные методы ZET и ZETM [11], главных компонент, «неподвижной точки» [12], заполнение безусловными и условными средними [10], локального заполнения [13]. В работе [14] предлагается способ совмещения процедуры кластеризации статистических данных и обработки пропущенных в них значений, основанный на модификации алгоритма конкурентного обучения сети Кохонена для проведения сферической кластеризации. В [15] представлена достаточно подробная классификация методов заполнения пропусков в данных.

Как правило, методы заполнения пропусков в данных основаны на различного рода эвристических процедурах, часто весьма сложных и имеющих итерационный характер, что в совокупности с известной трудностью во введении формализованной меры качества восстановления пропущенных элементов делает эти методы легко уязвимыми для обоснованной критики. К числу критикуемых относят (см., например, [13]) следующие недостатки "комплексирующих" процедур:

– вычисление параметров алгоритма заполнения пропусков по присутствующим данным, что вносит зависимость между наблюдениями;

– ухудшение качества оценок с ростом доли пропусков;

– невозможность проведения строгого исследования свойств алгоритма;

– искажение природы данных и характера выводов.

Представляется, что более оправданным является заполнение пропусков не числами, а интервалами, с привлечением экспертов. Последние, возможно, привлекая упомянутые выше и другие методы точечного заполнения, а также руководствуясь своими профессиональными знаниями об объекте исследования, могут доопределить исходную выборку, но не числами, а интервалами, преобразовав ее к виду $(\boldsymbol{Z}, h)$, где

$$z_{kt} = \begin{cases} x_{kt}, \text{если } a_{kt}^{x} - 1 \\ [x_{kt}^{-}, x_{kt}^{+}], \text{если } a_{kt}^{x} = 0, \end{cases}$$

$$h_{k} = \begin{cases} y_{k}, \text{если } b_{k}^{y} = 1 \\ [y_{k}^{-}, y_{k}^{+}], \text{если } b_{k}^{y} = 0. \end{cases}$$

Безусловно, для экспертов заполнять пропуски в данных интервалами проще и менее ответственно, чем числами, тем более всегда при возникновении каких-либо сомнений можно «раздвинуть» концы интервалов. При этом надо иметь в виду, что от экспертов не следует требовать высказывания любых соображений вероятностного или иного характера, уточняющих расположение «истинных» значений пропущенных элементов внутри или на границах указанных интервалов.

Проблемы обработки данных с интервальной неопределенностью изучались во многих работах (см., в частности [1,16-19] и краткий обзор, данный в [18]). При этом в [16-18] рассматривается метод распознающего функционала, в работах же [1,19] описаны способы построения линейной регрессии (2), основанные на точечной характеризации множеств решений интервальной системы линейных алгебраических уравнений (ИСЛАУ).

Такая характеризация по отношению к поставленной задаче может быть произведена следующим образом. Введем в рассмотрение матрицы Z^{-}, Z^{+} и вектора y^{-}, y^{+}:

$$Z^{-} = \left\| z_{ki}^{-} \right\|,\ Z^{+} = \left\| z_{ki}^{+} \right\|,\ h^{-} = (h_{1}^{-}, ..., h)^{T},$$

$$y^{+} = (h_{1}^{+}, ..., h_{n}^{+})^{T},\ k = \overline{1,n},\ i = \overline{1,m},$$

где

$$z_{kt}^{-} = \begin{cases} x_{kt}, \text{если } a_{kt}^{x} = 1 \\ x_{kt}^{-}, \text{если } a_{kt}^{x} = 0, \end{cases}$$

$$z_{kt}^{+} = \begin{cases} x_{kt}, \text{если } a_{kt}^{x} = 1 \\ x_{kt}^{+}, \text{если } a_{kt}^{x} = 0, \end{cases}$$

$$h_{k}^{-} = \begin{cases} y_{k}, \text{если } b_{k}^{y} = 1 \\ y_{k}^{-}, \text{если } b_{k}^{y} = 0, \end{cases}$$

$$h_{k}^{+} = \begin{cases} y_{k}, \text{если } b_{k}^{y} = 1 \\ y_{k}^{+}, \text{если } b_{k}^{y} = 0. \end{cases}$$

Тогда задача построения линейной регрессии (2) на основе сформированной с помощью экспертов выборки $(\boldsymbol{Z}, h)$ представима в виде задачи точечной характеризации множеств решений ИСЛАУ:

$$\left[Z^{-}, Z^{+} \right] a = \left[h^{-}, h^{+} \right] \qquad (3)$$

Рассмотрим возможные способы ее решения, используя описанный в [19] подход.

2. Определение оценок параметров модели (1) при различном характере исходной неопределенности в данных

Неопределенность в исходной информации, состоящая в присутствии в ней пропусков, может быть трех типов:

– пропуски связаны только с независимыми переменными, то есть $b^{y} = (1, ..., 1)$;

– пропуски связаны только с зависимой переменной, то есть $a_{ki}^{x} = 1,\ k = \overline{1,n},\ i = \overline{1,m}$;

– пропуски связаны и с зависимой, и с независимыми переменными, то есть существуют пары номеров (*k*, *i*), для которых $a_{ki}^{x}=0$ и (или) $b_{k}^{y}=0$.

Для каждого из них вектор параметров (α^1, α^2, или α^3) линейной регрессии (2) рассчитывается посредством решения задач линейного или частично-булевого линейного программирования (ЧБЛП), к которым сводится поиск так называемых квазирешений ИСЛАУ (3) [19], состоящий во внесении «искажений» в неравенства, задающие множества решений ИСЛАУ, гарантирующих их справедливость, и последующей минимизации этих искажений.

a) $b^{y}=(1,...,1)$.

Искомый вектор оценок параметров α^1 является результатом решения следующей задачи ЧБЛП:

$$Z^{-}B-Z^{+}\gamma-u\le h^{-},$$
$$Z^{+}B-Z^{-}\gamma+v\ge h^{+},$$
$$0\le\beta_i\le\sigma_i B,$$
$$0\le\gamma_i\le(1-\sigma_i)B,$$
$$\sigma_i\in\{0,1\},\ i=\overline{1,m},$$
$$a^1=\beta-\gamma,$$
$$u\ge 0, v\ge 0,$$
$$\sum_{k=1}^{n}(u_k+v_k)\to\min.$$

где B – большое положительное число, $h^{-}=h^{+}$.

b) $a_{ki}^{x}=1,\ k=\overline{1,n},\ i=\overline{1,m}$.

Вектор оценка параметров a^2 для данного случая является результатом решения задачи ЛП:

$$Z^{-}B-Z^{+}\gamma+u\ge h^{-},$$
$$Z^{+}B-Z^{-}\gamma-v\ge h^{+},$$
$$a^2=\beta-\gamma,$$
$$u\ge 0, v\ge 0,\ \beta\ge 0, \gamma\ge 0$$
$$\sum_{k=1}^{n}(u_k+v_k)\to\min$$

При этом $Z^{-}=Z^{+}$.

c). Существуют пары номеров (*k*, *i*), для которых $a_{ki}^{x}=0$ и (или) $b_{k}^{y}=0$.

Наконец, вектор оценок параметров a^3 для данного случая также является результатом решения задачи ЧБЛП:

$$Z^{-}B-Z^{+}\gamma-u\ge h^{+},$$
$$Z^{+}B-Z^{-}\gamma+v\ge h^{-},$$
$$0\le\beta_i\le\sigma_i B,$$
$$0\le\gamma_i\le(1-\sigma_i)B,$$
$$\sigma_i\in\{0,1\},\ i=\overline{1,m}$$
$$a^3=\beta-\gamma,$$
$$u\ge 0, v\ge 0,$$
$$\sum_{k=1}^{n}(u_k+v_k)\to\min.$$

При построении экспертно-статистических моделей по неполным данным на практике возможны редкие ситуации, когда одна из сформулированных задач ЛП или ЧБЛП будет иметь множество решений (или, что то же, соответствующее множество решений ИСЛАУ (3) окажется непустым). Пусть таковой оказалась последняя задача для определения вектора α^3. Формально возможная множественность ее решения может состоять в равенствах u=v=0. В этом случае, как рекомендовано в [19], может быть использован прием, применяемый в теории принятия решений [20] и состоящий в максимизации разрешающей способности задающих множества решений ИСЛАУ неравенств. Реализация этого приема приведет к преобразованию последней задачи в следующую:

$$Z^{-}B-Z^{+}\gamma+u\le h^{+},$$
$$Z^{+}B-Z^{-}\gamma-v\le h^{-},$$
$$0\le\gamma_i\le\sigma_i B,$$
$$0\le z_i\le(1-\sigma_i)B,$$
$$a^3=\beta-\gamma,$$
$$u\ge 0, v\ge 0,$$
$$\sigma_i\in\{0,1\},\ i=\overline{1,m},$$
$$\sum_{k=1}^{n}(u_k+v_k)\to\max.$$

Аналогичным образом в этом случае будут трансформированы и задачи для определения α^1 и a^2.

Следует обратить внимание на то, что только задача для определения вектора оценок α^2 имеет линейный вид. Задачи поиска векторов α^1 и a^3 включают в свой состав наряду с вещественными еще и булевы переменные $\sigma_i, i=\overline{1,n}$

Отметим одно важное обстоятельство. Если выборка (X,y) задана обычным для регрессионного анализа образом, т.е. все ее элементы заданы числами, имеют место равенства $a^1=a^2=a^3=a^{MHM}$, где a^{MHM} – оценка параметров уравнения (2), полученная с помощью метода наименьших модулей.

3. Моделирование объема импорта крупнотоннажных контейнеров

В настоящем разделе предлагаемый выше подход к построению ЭСМ применен для моделирования перевозок крупнотоннажных контейнеров на железнодорожном пункте пропуска Забайкальск-Маньчжурия с использованием статистической информации, представленной в [21]. Около 70% всего внешнеторгового грузопотока между КНР и РФ в сухопутном сообщении обеспечивает именно этот пункт пропуска, который расположен на железнодорожной грузовой межгосударственной передаточной станции Забайкальск.

Введем обозначения:

y – объем импорта крупнотоннажных контейнеров, ДФЭ (в 20-футовом эквиваленте);

x_1 – объем экспорта крупнотоннажных контейнеров, ДФЭ;

x_2 – грузооборот железнодорожного транспорта КНР, млрд. т-км.

Статистическая информация за 2006-2015 гг. по этим переменным представлена в таблице 1.

Таблица 1

Статистические данные

Номер	y	x_1	x_2
1	2168	9445	2195,4
2	4487	15597	2379,7
3	4677	15564	2510,6
4	1980	10675	2523,9
5	4109	9825	2764,9
6	5157	15169	2946,6
7	5255	36257	2918,7
8	4682	22442	2917,4
9	4496	25681	2753
10	443	23011	2342,1

Построим на этих данных уравнение

$$y = a_0 + a_1 x_1 + a_2 x_2 \qquad (4)$$

При этом вначале воспользуемся традиционными для регрессионного анализа методами наименьших квадратов (МНК) и модулей (см., например, [1]), а затем, в предположении присутствия пропусков в данных, представленных в таблице 1, определим параметры (4) с помощью предложенного в работе подхода.

После применения МНК регрессия (4) принимает вид:

$$y = -14285.9 + 0.597 x_1 + 17.054 x_2, \qquad (5)$$

Использование МНМ приводит к уравнению

$$y = 23688.9 + 0.405 x_1 + 4.854 x_2, \qquad (6)$$

$$M = 54907.48.$$

Здесь M – сумма модулей ошибок аппроксимации.

Заметное различие МНК- и МНМ-оценок в регрессиях (5) и (6) может быть вызвано присутствием явного выброса (наблюдения, не согласующего со всей выборкой в целом) в четвертом наблюдении, а, как известно, МНК и МНМ по-разному реагируют на подобные ситуации. Так, МНМ, по существу, выбросы игнорирует.

Предположим теперь, что исходная выборка содержит четыре пропуска в данных для независимых переменных, а именно:

$$a_{21}^x = 0, a_{52}^x = 0, a_{71}^x = 0, a_{82}^x = 0.$$

Таким образом, определенность в данных в этом случае составит $\theta = 0.87$.

Пусть эксперты, используя упомянутые выше способы заполнения пропусков в данных, а также привлекая свои знания и опыт, заполнили пропуски интервалами:

$$z_{21} = [15550, 15650], z_{52} = [2700, 2800],$$
$$z_{71} = [36000, 36800], z_{82} = [2850, 2950].$$

В этом случае вектор оценок параметров уравнения (4) после решения соответствующей задачи ЧБЛП примет вид:

$$a^1 = (22979, 0.396, 5.251), \; M^1 = 54626.97,$$

где M^1 – значение ее целевой функции на оптимальном решении, являющееся аналогом величины M.

Таким образом, замена лишь четырех элементов выборки из тридцати возможных на весьма узкие интервалы, содержащие эти элементы, привела к некоторому изменению оценок параметров уравнения (4) по сравнению с МНМ-оценкой при малом – на 0.51%, – уменьшении суммарной ошибки аппроксимации.

Заключение

В работе решается проблема построения линейного регрессионного уравнения по исходной информации, содержащей пропуски, с использованием как статистической, так и экспертной информации. Развиваемый подход предполагает заполнение пропусков интервалами, формируемыми экспертами, которые руководствуются при этом как своими опытом и знаниями, так и используют методы заполнения пропусков числами. После этого оценивание параметров модели в зависимости от характера исходной неопределенности в данных сводится к решению задач линейного или частично-булевого линейного программирования.

Решена задача построения линейного регрессионного уравнения влияния объема экспорта крупнотоннажных контейнеров и грузооборота железнодорожного транспорта КНР на объем импорта крупнотоннажных контейнеров на железнодорожном пункте пропуска Забайкальск-Маньчжурия. При этом было установлено, что даже незначительная исходная неопределенность в данных может вызвать существенное изменение численных оценок параметров модели.

Литература

1. *Носков С.И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: Облинформпечать, 1996. 320 с.
2. *Носков С.И., Торопов В.*Д. Формирование исходной информации и идентификация параметров экспертной модели статистического типа // Современные технологии. Системный анализ. Моделирование. 2005. №4. С.103-106.
3. *Головченко В.Б., Носков С.И.* Выбор класса линейной по параметрам регрессии на основе экспертных высказываний // Кибернетика и системный анализ. 1992. №5. С.109-115.
4. *Головченко В.Б., Носков С.И.* Комбинирование прогнозов с учетом экспертной информации // Автоматика и телемеханика. 1992. №11. С.109-117.
5. *Головченко В.Б., Носков С.И.* Оценивание параметров эконометрической модели по статистической и экспертной информации // Автоматика и телемеханика. 1991. №4. С.123-132.
6. *Головченко В.Б., Носков С.И.* Прогнозирование на основе дискретной динамической модели с использованием экспертной информации // Автоматика и телемеханика. 1993. №10. С.140-148.
7. *Мандель А.С.* Метод аналогов в прогнозировании коротких временных рядов: экспертно-статистический подход // Автоматика и телемеханика. 2004. № 4. С. 143-152.
8. *Лисицын Д.В.* Комбинированные регрессионные модели для описания данных, представленных в разных шкалах // Сборник научных трудов Новосибирского государственного технического университета. 2013. № 3 (73). С. 41-48.

9. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Финансы и статистика,1981. т.1. 366 с., т.2. 351с.

10. *Литтл Р.Дж., Рубин Д.Б.* Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991. 334с.

11. *Загоруйко Н.Г., Елкина В.Н., Темиркаев В.С.* Алгоритм ZET–75 заполнения пробелов в эмпирических таблицах и его применение // Машинные метода обнаружения закономерностей. Новосибирск: Наука, 1976. С.57-63.

12. *Айвазян С.А., Енюков И.С, Мешалкий Л.Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. 472 с.

13. *Никифоров А.М.* Разработка и исследование статистических методов распознавания образов с самообучение и обработка неполных данных. Дис… канд. физ.-мат.наук. М.1987. 144с.

14. *Ефимов А.С.* Решение задачи кластеризации методом конкурентного обучения при неполных статистических данных // Вестник Нижегородского университета им. Н.И. Лобачевского. 2010. № 1. С. 220-225.

15. *Рыженкова К.В.* Методы восстановления пропуска данных при проведении статистических исследований // Интеллект. Инновации. Инвестиции. 2012. № 3. С. 127-133.

16. *Шарый С.П.* Задача восстановления зависимостей по данным с интервальной неопределенностью // Заводская лаборатория. Диагностика материалов. 2019. Т.86. №1. С.62-74.

17. *Шарый С.П., Шарая И.А.* Распознавание разрешимости интервальных уравнений и его приложения к анализу данных // Вычислительные технологии. 2013. Т. 18. №3. С.80-109.

18. *Shary S.P.* Maximum consistency method for data fitting under interval uncertainty // J. of Global Optimization. 2015. Vol.62. No.3. 16 p.

19. *Носков С. И.* Точечная характеризация множеств решений интервальных систем линейных алгебраических уравнений // Информационные технологии и математическое моделирование в управлении сложными системами. 2018. № 1. С. 8-13.

20. *Васильев С.Н., Селедкин А.П.* Синтез функции эффективности в многокритериальных задачах принятия решений // Известия АН СССР. Тех. кибернетика. 1980. №3. С.186-190.

21. *Носков С.И., Базилевский М.П.* Построение регрессионных моделей с использованием аппарата линейно-булевого программирования. Иркутск, 2018. 176 с.

CONSTRUCTION OF EXPERT-STATISTICAL MODELS FROM INCOMPLETE DATA

Sergey I. Noskov, *Irkutsk State Transport University, Irkutsk, Russia,* ***sergey.noskov.57@mail.ru***

Abstract

The article deals with the problem of constructing a linear regression model based on incomplete data containing gaps, using statistical and expert information. The reasons for the gaps in the data can be, in particular, a temporary malfunction (failure) of the measuring equipment when taking various technical characteristics, or negligence in the work of statistical services when fixing the reporting indicators. Very often, gaps arise when processing various kinds of sociological information in the form of questionnaires, when respondents refuse to answer a specific question (but answer others) or give an inadmissible, in particular, evasive answer. The approach proposed in the work involves filling the gaps with intervals, the boundaries of which are formed by experts, guided by both their experience and knowledge about the object of research, and using the well-known methods of point filling in the gaps. After that, the estimation of the parameters of the model, depending on the nature of the initial uncertainty in the data, is reduced to solving problems of linear or partially Boolean linear programming. The case is considered when the solution of the formalizing uncertainty in the initial data of the interval system of linear algebraic equations is not unique. The problem of constructing a linear regression equation for the influence of the volume of export of large-tonnage containers and the freight turnover of the PRC railway transport on the volume of import of large-capacity containers at the Zabaikalsk-Manchuria railway checkpoint is solved.

Keywords: *regression model, data uncertainty, gaps, interval system of linear algebraic equations, quasi-solutions, parameter estimates.*

References

1. S.I. Noskov (1996). A technology for modeling objects with unstable functioning and uncertainty in data. Irkutsk: Oblinformpechat. 320 p.
2. S.I. Noskov, V.D. Toropov (2005). Formation of initial information and identification of parameters of an expert model of statistical type. *Modern technologies. System analysis. Modeling.* No. 4. C.103-106.
3. V.B. Golovchenko, S.I. Noskov (1992). The choice of a class of regression linear in parameters based on expert statements. *Cybernetics and Systems Analysis.* No. 5. P. 109-115.
4. V.B. Golovchenko, S.I. Noskov (1992). Combining forecasts taking into account expert information. *Automation and Telemechanics.* No. 11. P. 109-117.
5. V.B. Golovchenko, S.I. Noskov (1991). Estimation of the parameters of an econometric model based on statistical and expert information. *Automation and Telemechanics.* No. 4. P. 123-132.
6. V.B. Golovchenko, S.I. (1993). Noskov Prediction based on a discrete dynamic model using expert information. *Automation and Telemechanics.* No. 10. P. 140-148.
7. A.S. Mandel (2004). Method of analogues in forecasting short time series: an expert-statistical approach. *Automation and telemechanics.* No. 4. P. 143-152.
8. D.V. Lisitsyn (2013). Combined regression models for describing data presented in different scales. *Collection of scientific papers of the Novosibirsk State Technical University.* No. 3 (73). P. 41-48.
9. N. Draper, G. Smith (1981). Applied regression analysis. Moscow: Finance and Statistics. Vol. 1. 366 p., Vol. 2. 351p.
10. R.J. Little, D.B. Rubin (1991). Statistical analysis of data with gaps. Moscow: Finance and Statistics. 334 p.
11. N.G. Zagoruiko, V.N. Elkina, V.S. Temirkaev (1976). Algorithm ZET - 75 for filling the gaps in empirical tables and its application. *Machine methods for detecting patterns.* Novosibirsk: Nauka. P. 57-63.
12. S.A. Ayvazyan, I.S. Enyukov, L.D. Meshalky (1983). Applied statistics. Basics of modeling and primary data processing. Moscow: Finance and Statistics. 472 p.
13. A.M. Nikiforov (1987). Development and research of statistical methods for pattern recognition with self-learning and processing of incomplete data. Dis ... Cand. physical and mathematical sciences. Moscow. 144 p.
14. A.S. Efimov (2010). Solution of the clustering problem by the method of competitive learning with incomplete statistical data. *Bulletin of Nizhny Novgorod University. N.I. Lobachevsky.* No. 1. P. 220-225.
15. K.V. Ryzhenkova (2012). Methods for restoring missing data in statistical studies. *Intellect. Innovation. Investments.* No. 3. P. 127-133.
16. S.P. Shary (2019). The problem of recovering dependencies from data with interval uncertainty. *Zavodskaya laboratory. Diagnostics of materials.* Vol. 86. No. 1. P. 62-74.
17. S.P. Shary, I.A. Sharaya (2013). Recognition of the solvability of interval equations and its applications to data analysis. *Computational technologies.* Vol. 18. No. 3. P. 80-109.
18. S.P. Shary (2015). Maximum consistency method for data fitting under interval uncertainty. *J. of Global Optimization.* Vol. 62. No. 3. 16 p.
19. S.I. Noskov (2018). Point characterization of solution sets of interval systems of linear algebraic equations. *Information technologies and mathematical modeling in control of complex systems.* No. 1. P. 8-13.
20. S.N. Vasiliev, A.P. Seledkin (1980). Synthesis of the efficiency function in multicriteria decision-making problems. *Izvestiya AN SSSR. Those. cybernetics.* No. 3. P. 186-190.
21. S.I. Noskov, M.P. Bazilevsky (2018). Construction of regression models using the apparatus of linear-Boolean programming. Irkutsk. 176 p.

Information about author:

Sergei I. Noskov, *Doctor of Technical Sciences, Professor of the Department of Information Systems and Information Security, Irkutsk State University of Railways, Irkutsk, Russia*