УДК 004.6

Мартышенко Сергей Николаевич

Владивостокский государственный университет экономики и сервиса
Владивосток, Россия

Методы восстановления пропусков в данных, представленных в различных измерительных шкалах

В работе предложен метод восстановления многомерных данных, полученных в ходе социально-экономических исследований. Основное преимущество метода заключается в использовании восстанавливающих признаков различной природы, что существенно расширяет диапазон применения метода. Предложены методы оценки качества восстановления, основанные на использовании процедуры скользящего экзамена. Рассмотренный метод реализован программно и прошел апробацию на модельных и реальных данных.

Ключевые слова и словосочетания: данные с пропусками, методы восстановления данных, многомерный статистический анализ, качественные данные, моделирование данных.

Большинство исследователей, которые проводят исследования социально-экономических процессов, сталкиваются с проблемой пропуска данных или неответа в таблицах объект-свойство [12]. Иначе еще эту проблему называют проблемой неполноты данных [3]. Часто выбросы тоже можно рассматривать как пропущенные данные. К выбросам можно отнести данные, которые явно противоречат данным всей выборки. Причем, противоречие может возникать не только со значениями одного признак, но и со значениям прочих признаков одного наблюдения. В обоих случаях перед исследователем стоит дилемма: либо отбросить всю строку таблицы данных? либо каким-то образом исправить ошибку (восстановить данные). Часть противоречий (ошибок) может быть выявлена и исправлена на предварительных этапах анализа данных путем логического анализа противоречий в многомерных данных. Для этого можно использовать специальные средства [9].

При большом количестве исследуемых признаков количество пропусков может быть значительным. Часто отбрасывать данные нежелательно по той причине, что на основании многомерных данных решается множество задач, в которых используются либо одномерные признаки (частотные ряды), либо часть признаков многомерных наблюдений. В одной задаче все признаки задействуются крайне редко. Если го-

ворить об анкетных данных, то анкеты могут включать много вопросов, которые служат для классификации данных (например, данные по социально-демографическому портрету респондентов), а следовательно задействованы при решении определенного круга задач.

Многообразие ситуаций и причин возникновения пропусков в данных породило множество исследований в этой области. Особенно много работ, посвященных исследованию данной проблемы, в зарубежных источниках. Обширный список таких работ можно найти в отдельных работах отечественных ученых [2, 14]. Большое количество методов потребовало систематизации подходов и разработки классификации методов [1, 3, 4, 5]. Многие авторы за основу принимают схему классификации, представленную в работах [13, 15] (рис. 1). В указанных выше работах приводятся основные принципы распространенных методов восстановления данных. Можно отметить, что новые разрабатываемые методы, как правило, вписываются в представленную схему классификации.

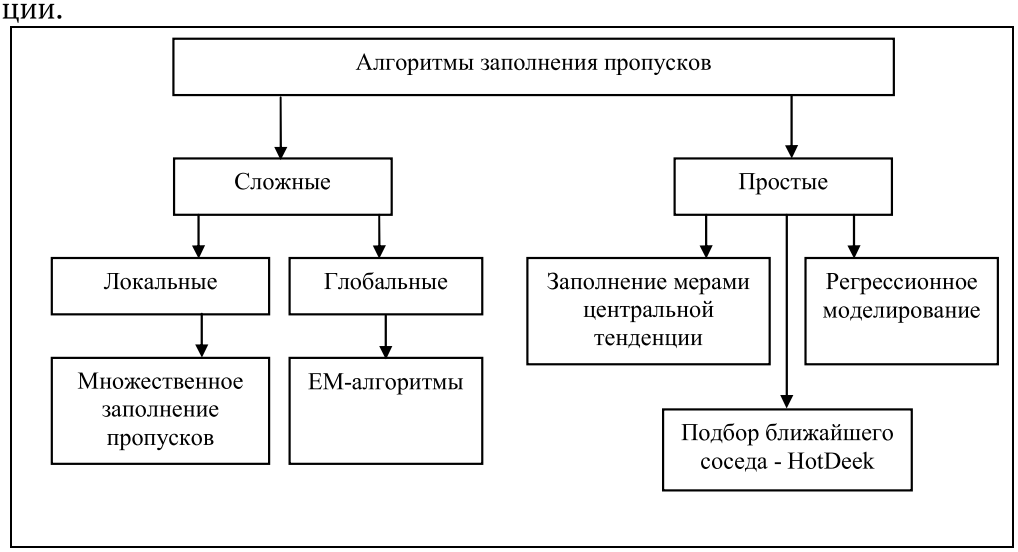

Рис. 1. Классификация методов заполнения пропусков

Можно утверждать, что теория восстановления пропущенных данных постоянно развивается, и соответственно появляются новые алгоритмы и модернизируются известные. Это связано с тем, что не может быть разработано абсолютного алгоритма, который был бы применим и давал наилучшие результаты во всех ситуациях. Многие исследователи, доказывая преимущество того или иного подхода или метода, демонстрируют достоинства метода на конкретном примере. Но примеры тоже является частными случаями и не доказывают полного превосходства одного метода над другим. Несмотря на существование большого коли-

чества методов восстановления данных в широко известных пакетах по обработке статистических данных представлены только простейшие алгоритмы, которые во многих случаях не дают требуемой точности. То есть, задача восстановления данных сейчас во многом носит исследовательский характер и используется специалистами, более менее представляющими механизм работы используемых алгоритмов. Сохраняется как теоретическая проблема оценки точности результатов, полученных в результате применения алгоритмов восстановления.

В данной работе предлагается к рассмотрению метод восстановления данных, который может использоваться в ситуации, когда большинство известных методов не применимо. В большинстве методов восстановления данных используются признаки, измеренные в шкале отношений. При исследовании социально-экономических процессов часто получают данные, представленные в различных шкалах. Мы предлагаем алгоритм, которые позволяют работать с различными признаками. Конечно, предложенный алгоритм тоже не всегда гарантирует получения требуемой точности. Возможности алгоритма всегда ограничены имеющимися данными и их латентной структурой. В дополнение к алгоритму предлагается несколько процедур оценки точности результатов, что позволяет исследователю самому принять решение о приемлемости полученного результата.

Алгоритм, основанный на разработке эталонов классифицированных данных

Алгоритм основан на предположении случайности возникновения пропусков данных в таблице объект-свойство. Для такого предположения часто используется аббревиатура MCAR (missing completely at random). Это предположение принимается в большинстве известных алгоритмов. Чаще всего предположение выполняется и его можно проверить с помощью известных статистических методов. В таблице данных допускается присутствие данных, измеренных в различных шкалах. Таблицу данных представим в форме отсортированной таблицы (рис. 2). Таблица содержит m+1 столбец. Первые m столбцов ($X_1 \ X_2 \ \dots \ X_i \ \dots \ X_m$) содержат значения признаков, не имеющих пропусков. Эти признаки будем называть восстанавливающими признаками. Столбец Y содержит признак, в котором допущены пропуски. Этот признак будем называть восстанавливаемым признаком. Первые n_0 содержат наблюдения без пропусков. Следующие n_1 строк имеют пропуски в признаке Y. То есть, необходимо восстановить n_1 значений признака Y.

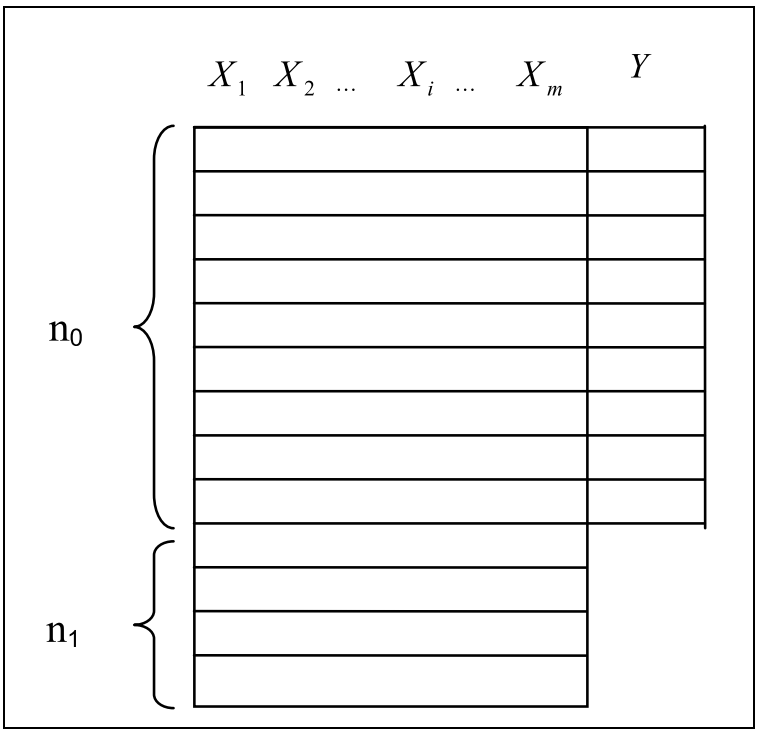

Рис. 2. Представление таблицы данных

Процедура более эффективно работает при восстановлении числовых признаков, но при достаточно большом количестве данных (не менее тысячи) можно пытаться восстанавливать и данные других типов. Для простоты будем считать, что данные числовые. Рассмотрим работу алгоритма по этапам.

Первый этап. Осуществляется преобразование всех числовых значений признаков к ранговым значениям (операция ранжирования). Признаки номинальные и ранговые не преобразуются. При этом номинальные признаки должны иметь небольшое количество значений (желательно меньше 10). Иначе номинальные признаки нужно подвергнуть предварительной обработке, приводя их к структурированному виду. Для этого применяются процедуры обработки качественных данных, описанные в работе [5].

Процедура ранжирования заключается в разбиении значений признака на равные интервалы и замене исходных значений ранговыми (номерами интервалов). Количество интервалов *r* должно быть не очень большим (рекомендуется 5), иначе могут появиться интервалы без значений, что нежелательно (но ситуация допустимая). Ранговые признаки ранжировать нет необходимости, и можно использовать имеющуюся

систему рангов. Ранжированные значения обозначим той же буквой только со штрихом.

Далее выборка (таблица данных) разбивается на две части, которые далее рассматриваются по отдельности. Первую выборку можно назвать «обучающей выборкой», вторую «контрольной выборкой».

Второй этап. Производится сортировка «обучающей выборки» по рангам признака Y'. Пусть признак Y имеет k рангов (классов). Отсортированная выборка представлена на рис. 3. Сумма количества наблюдений по классам равна объему «обучающей выборки» n_0 (1).

$$n_0 = s_1 + s_2 + s_3 + \cdots + s_k \tag{1}$$

Заметим, что в таблице на рис. 3 столбец Y' содержит k групп повторяющихся значений.

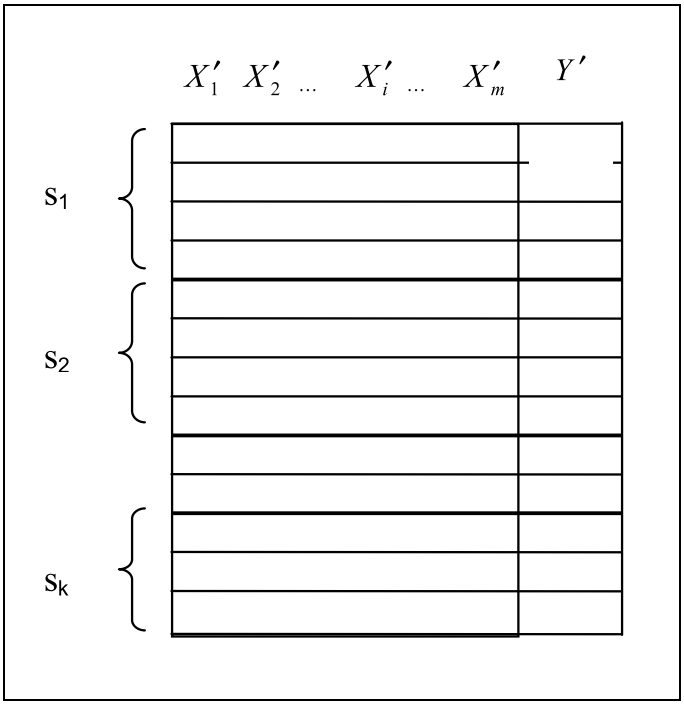

Рис. 3. Классифицированная таблица ранговых значений признаков

Третий этап. По данным каждого столбца X'_j рассчитывается таблицы абсолютных условных частотных рядов признаков по классам. Каждая такая таблица будет содержать k строк (по количеству классов) и r столбцов (по количеству градаций признаков X'_j. После этого таблицы нормируются по строкам путем деления на соответствующее количество элементов класса $s_t (t = \overline{1,k})$. Тогда сумма элементов строк каждой таб-

лицы будет равна единице. Нормированные таблицы представлены на рис. 4. Эти частотные ряды представляют собой выборочные условные распределения переменных X при заданных значениях Y.

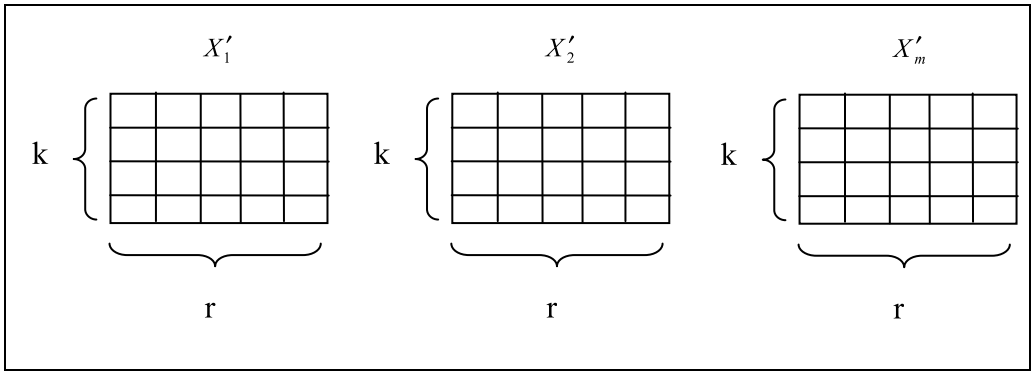

Рис. 4. Частотные ряды признаков по классам значений признаков

Четвертый этап. На этом этапе рассчитывается вектор строка «эталонов» классов выборки. Эталон состоит из m частей по количеству признаков X. Каждая часть эталона состоит из r разрядов по количеству дискретных значений признаков X'. Макет эталона представлен на рис. 5. Рассмотрим правило расчета элементов эталона. Каждая часть эталона рассчитывается по соответствующей таблице, представленной на рис. 4. Общее количество столбцов во всех таблицах также равна $r \times m$. Соответственно размерность эталона тоже $r \times m$. Для расчета каждого элемента используются данные одного столбца таблицы. По данным каждого столбца определяется максимальное значение и номер строки (номер класса) присваивается соответствующему, элементу эталона.

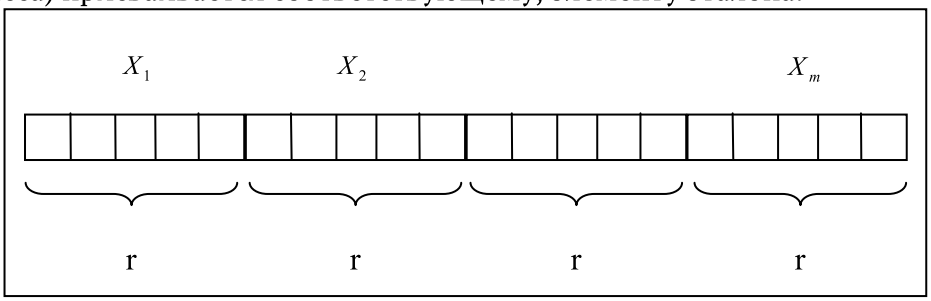

Рис. 5. Макет эталона классов

Процедуре расчета элементов эталона можно дать геометрическую интерпретацию. На рис. 6 представлена графическая интерпретация расчета одной части эталона. Все остальные части рассчитываются аналогично. При расчете пятого элемента эталона для примера, приведенного на рис. 6, возникает неопределенная ситуация. Неопределенность состоит в том, что максимум достигается сразу в двух строках –

второй и третьей. В этом случае предпочтение отдается тому классу (строке таблицы условных распределений), в котором количество элементов класса s_t больше. Предположим, что в нашем случае $s_5 > s_2$.

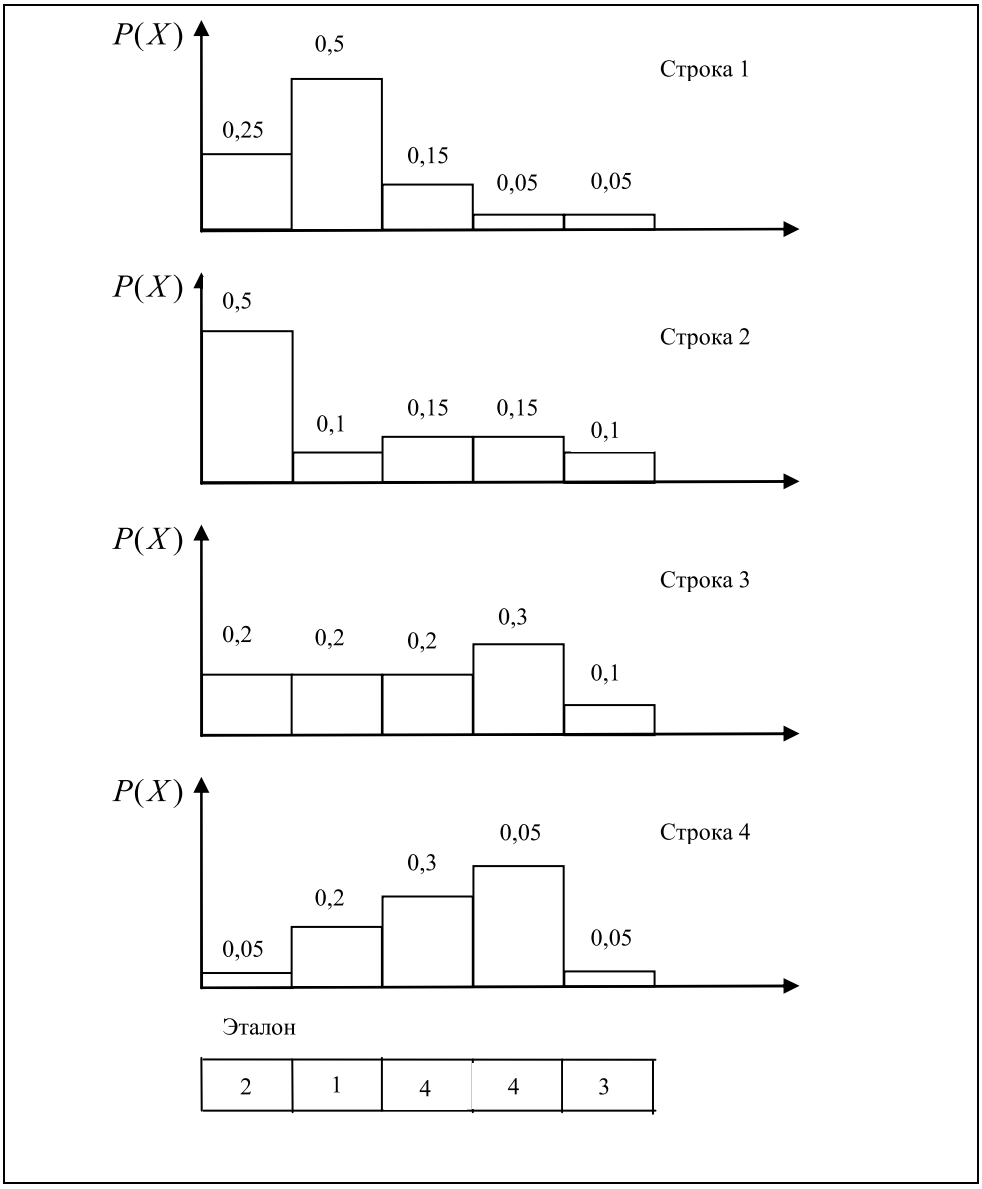

Рис. 6. Графическая интерпретация расчета одной части эталона классов

Пятый этап. На этом этапе производится сравнение многомерных данных «контрольной выборки» с эталоном и прогнозирование номера класса восстанавливаемого признака Y для наблюдений «контрольной

выборки». Процедуру сравнения продемонстрируем на числовом примере (рис. 7).

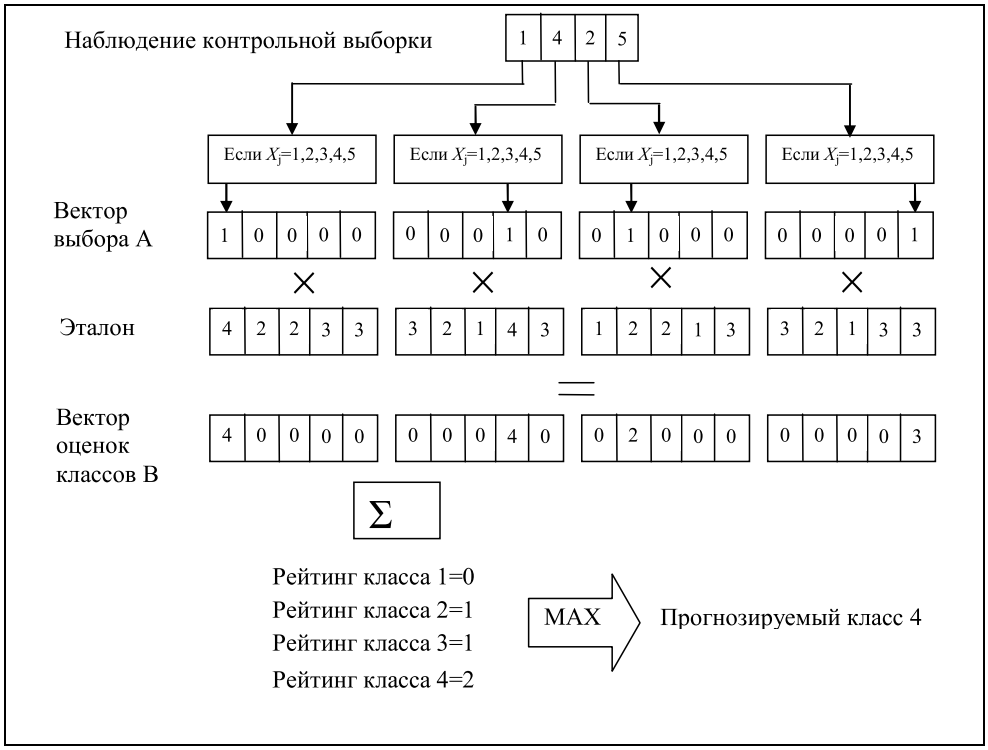

Рис. 7. Процедура сравнения наблюдения «контрольной выборки» с эталоном

Расчет производится за несколько шагов:

1. Формируется вспомогательный вектор выбора А. Значения «1» указывают номер интервала в котором лежит значение признака X_j;

2. Формируется вспомогательный вектор В, как произведение элементов вектора А и вектора эталона;

3. Подсчитываются рейтинги классов, как количество оценок по каждому классу в векторе оценок классов В;

4. Определяется прогноз класса для наблюдения «контрольной выборки» по максимальному рейтингу класса.

На четвертом шаге опять может возникнуть неопределенность. Она возникает, когда максимум рейтинга класса достигается сразу для нескольких классов. В этом случае предпочтение тоже отдается классу с наибольшим объемом выборки $s_t(t = \overline{1,k})$.

Шестой этап.

Это заключительный этап. На этом этапе спрогнозированные значения номеров классов для элементов «обучающей выборки» заме-

няются средними значениями средних значений признака *Y*, рассчитанными по данным «обучающей выборке».

Рассмотренный алгоритм в соответствии с классификацией, представленной на рис. 1, можно отнести к классу сложных, глобальных. Этот алгоритм относится к классу сложных не в силу сложности расчетов и множества этапов расчета, а в силу того, что при использовании алгоритма при решении конкретной задачи перед исследователем стоит проблема выбора. Необходимо задать количество интервалов для восстанавливающих признаков и восстанавливаемого признака. Возможно, для этого придется провести небольшой эксперимент.

В сложных алгоритмах исследователь должен хорошо представлять принцип работы алгоритма. Допуская возможность эксперимента, для оптимизации точности работы алгоритма, необходимо иметь критерии для сравнения различных вариантов построенного решающего правила.

Рассмотренный метод не столь чувствителен к подбору восстанавливающих признаков. Но проблема с подбором восстанавливающих признаков все-таки существует, потому что излишние «неинформативные» признаки могут «засорять» полезную информацию. Рекомендуется на первых этапах использовать не очень большое количество восстанавливающих признаков, постепенно наращивая их количество.

При подборе числовых признаков целесообразно включать сначала признаки с большей корреляцией с восстанавливаемым признаком. В случае использования ранговых признаков можно использовать ранговые коэффициенты корреляции. Не нужно исключать здравый смысл содержательного анализа признаков.

При восстановлении данных, полученных в ходе социально-экономических исследований, может оказаться очень полезным, в качестве восстанавливающего рангового признака включать некоторый обобщающий признак, сформированный на основании представлений исследователя о социально-демографическом профиле групп населения. Такой признак формируется на основе нескольких признаков.

Можно привести показательный пример недоучета факторов социологического портрета. Например, при восстановлении признака возраст может оказаться, что для наблюдения восстанавливаемого признака подходит возраст 70 и старше лет. И все-бы хорошо, и алгоритм сработал корректно, и средние ошибки минимальны, но если принять во внимание, что это данные по студенту дневной формы обучения, то возникают вопросы по корректности такого восстановления.

Методика оценки точности восстановления данных

Для сравнения результатов по точности восстановлению данных, полученных с помощью различных методов или различных данных, используемых для восстановления необходимы некоторые критерии качества.

Многие авторы считают, что после восстановления данных должны сохраняться основные свойства выборки (оценки функций плотности, средние и дисперсии признаков). При невысоком проценте восстанавливаемых данных эти параметры практически не изменяются после включения в выборку восстановленных данных.

По мнению автора, наиболее универсальным средством сравнения результатов являются оценки ошибок, рассчитанные методом скользящего экзамена [7].

Суть метода заключается в том, что решающее правило восстановления данных проверяется на данных обучающей выборки, которая содержит и восстанавливающие признаки и восстанавливаемый признак в полном объеме (полные данные). Процедура скользящего экзамена состоит в том, что из обучающей выборки последовательно отбрасывается по одному наблюдению, которые потом восстанавливаются с помощью оставшихся наблюдений.

Ошибкой считается отнесение наблюдений восстанавливаемого признака к другим классам. Процедура повторяется n_0 раз. При достаточно больших объемах выборки (в тысячах наблюдений) это может занять достаточно много времени, но при использовании современной вычислительной техники это не проблема. В результате процедуры скользящего экзамена рассчитывается матрица ошибок восстановления:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}, \quad (2)$$

где p_{ij} - количество значений восстанавливаемого признака, из класса *i,* отнесенных при восстановлении к классу *j*.

Иначе говоря, количество правильных восстановлений будет равно сумме диагональных элементов матрицы:

$$Q = \sum_{t=1}^{k} p_{tt} \quad (3)$$

Сумма элементов по строке дает объем класса обучающей выборки:

$$Q_t = \sum_{l=1}^{k} p_{tl} = s_t, \quad t = \overline{1,k} \tag{3.1}$$

Качество восстановления можно оценить показателем Ω – процент ошибок восстановления и показателями Ω_t - процентом ошибок по классам:

$$\Omega = 1 - \frac{n_0 - Q}{n_0} \tag{4}$$

$$\Omega_t = 1 - \frac{p_{tt}}{s_t} \tag{5}$$

Более подробный анализ ошибок позволяет сделать нормированная матрица ошибок восстановления. Нормированные значений матрицы ошибок производится путем делением каждой строки на объем класса обучающей выборки $s_t (t = \overline{1,k})$. Тогда сумма элементов каждой строки матрицы будет равна единице. Элементы нормированной матрицы ошибок P' обозначим p'_{ij}:

$$P' = \begin{bmatrix} p'_{11} & p'_{12} & \cdots & p'_{1k} \\ p'_{21} & p'_{22} & \cdots & p'_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p'_{k1} & p'_{k2} & \cdots & p'_{kk} \end{bmatrix} \tag{6}$$

Ошибки по классам могут быть распределены неравномерно и возможно придется отказаться от восстановления некоторых данных. Матрицы ошибок восстановления могут расширить понимание причин возникновения неответов, что важно для организации последующих исследований при мониторинге социально-экономических процессов. Многие авторы считают проблему выявления причин неответов еще более важной, чем проблему восстановления данных.

Изложенная процедура оценки уровня ошибок восстановления пригодна для любого типа восстанавливаемых признаков. При восстановлении числовых признаков можно считать по обучающей выборке еще и дисперсии отклонений исходных значений и оценок, полученных при восстановлении данных. Такие оценки могут быть рассчитаны как по всей выборке, так и по классам.

Заключение

Прежде чем приступать к процедуре восстановления данных, необходимо провести тщательный анализ возможных ошибок в данных и выявить выбросы. Для этого мы используем процедуры, автоматизирующие данный процесс [6-8]. При больших объемах выборок и значительном количестве признаков без специальных программных средств не обойтись.

Необходимо заметить, что заниматься трудоемкой процедурой восстановления данных имеет смысл, если исследователь предполагает в своей работе использование многомерных статистических методов. Многие исследователи ограничиваются исследованием одномерных признаков, и поэтому восстанавливать пропуски в данных им не имеет смысла, потому что можно внести только искажения в результаты, полученные по полным данным.

Некоторые многомерные статистические методы могут быть реализованы программно с учетом наличия пропусков в части данных. В качестве простейшего примера можно привести расчет ковариационной матрицы по данным с пропусками.

При расчете ковариационной матрицы используются оценки средних, которые могут быть рассчитаны по каждому признаку в отдельности с учетом пропусков. Можно было привести и более сложные примеры.

Однако можно заметить, что учет пропусков программно может привести к существенному усложнению программы, а при их использовании должны быть оговорены обозначения отсутствия данных в таблицах данных. При этом для разных признаков (по типу) должны быть свои условные обозначения. Поэтому такой подход применяется в исключительных случаях.

Мы, например, использовали его на предварительных этапах обработки данных при обнаружении выбросов и грубых ошибок.

Относительно рассмотренного алгоритма в качестве выводов по статье можно добавить, что метод может дать более точные результаты по сравнению с некоторыми другими методами за счет расширения диапазона используемых признаков.

Серия экспериментов на модельных данных показала преимущества метода перед другими. В экспериментах мы широко использовали программу моделирования многомерных нормальных распределений [10]. В настоящее время мы продолжаем эксперименты на модельных данных для выявления условий и ограничений применения метода. До-

казано, что любой метод очень сильно зависит от имеющихся данных. Если они «плохие», то тут бессильны самые совершенные методы.

Еще можно добавить, что разработанные нами программные средства, либо содержат встроенные блоки, позволяющие проводить эксперименты, либо включают параметры, автоматизирующие экспериментальную работу, облегчая работу исследователя.

Разработанные программные средства прошли апробацию на больших массивах данных, собранных в исследованиях посвященных анализу проблем туризма и рекреации в Приморском крае [11, 12].

1. Абраменкова, И.В., Круглов, В.В. Методы восстановления пропусков в массивах данных // Программные продукты и системы. – 2005. – № 2. – С. 4.
2. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
3. Зангиева, И.К. Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: методология, методы, математическое моделирование. – 2011. – № 33. – С. 028-056
4. Злоба, Е., Яцкив, И. Статистические методы восстановления пропущенных данных//Computer Modelling & New Technologies. – 2002. – Vol. 6. – No.1. – P. 51-61.
5. Мартышенко, Н.С. Компьютерная технология обработки качественных данных опросов потребителей туристских услуг // Маркетинг и маркетинговые исследования. – 2011. – №3. – С. 184-192.
6. Мартышенко, Н.С. Методическое обеспечение анализа поведения потребителей на региональном туристском рынке // Вестник Тихоокеанского государственного экономического университета. – 2005. – №4. – С. 19-31.
7. Мартышенко, Н.С., Мартышенко, С.Н. Технологии повышения качества данных в анкетном опросе // Практический маркетинг. – 2008. – №1. – С. 8–13.
8. Мартышенко, С.Н., Мартышенко, Н.С. Метод обнаружения ошибок в эмпирических данных // Известия вузов. Северо-Кавказский регион – 2008. – №1. – С. 11–14.
9. Мартышенко, С.Н., Мартышенко, Н.С., Кустов, Д.А. Многомерные статистические методы повышения достоверности маркетинговых данных // Практический маркетинг. – 2007. – №1. – С. 20–30.
10. Мартышенко, С.Н., Мартышенко Н.С., Кустов Д.А. Моделирование многомерных данных и компьютерный эксперимент // Техника и технология. – 2007. – №2. С. 47–52.

11. Мартышенко, Н.С. Принципы формирования туристского кластера в Приморском крае // Экономика региона. – 2009. – №1. – С. 204-208.
12. Мартышенко, Н.С. Вопросы анализа и прогнозирования пространственного развития рекреации и туризма // Регион: экономика, социология. – 2010. – №3. – С. 167-175.
13. Рыженкова, К.В. Методы восстановления пропуска данных при проведении статистических исследований // Интеллект. Инновации. Инвестиции. – 2012. – № 3. – С. 127-133.
14. Юдин, Г.Б. Территориальная локализация и уровень неответов в массовом опросе // Социологический журнал. – 2008. – № 1. – С. 49-72.
15. Kalton, G., Kasprzyk, D. The Treatment of Missing Survey Data // Survey Methodology. –1986. – No.12. – P. 1–16.