ОРГАНИЗАЦИЯ И МЕТОДЫ ОБРАБОТКИ ДАННЫХ СТАТИСТИЧЕСКОГО НАБЛЮДЕНИЯ

НОВЫЕ ПОДХОДЫ К ОРГАНИЗАЦИИ СТАТИСТИЧЕСКОГО НАБЛЮДЕНИЯ ЗА ДИФФЕРЕНЦИАЦИЕЙ ЗАРАБОТНОЙ ПЛАТЫ ПО ГРУППАМ ПРОФЕССИЙ И ДОЛЖНОСТЕЙ

Т.Л. Горбачева,
Л.А. Белоконная,
О.Б. Жихарева,
Федеральная служба государственной статистики

Основные направления статистического наблюдения за уровнем заработной платы работающих

Заработная плата является одним из важнейших индикаторов состояния экономики и уровня жизни населения. Развитие различных форм собственности, сокращение государственного сектора экономики и расширение прав предприятий в установлении систем заработной платы и решении вопросов, связанных с дополнительными расходами на содержание рабочей силы, привели к значительным изменениям в условиях оплаты труда работников различных отраслей экономики. Эти изменения проявились прежде всего в увеличении дифференциации работников по уровню заработной платы и перераспределении отраслей экономики по размеру среднемесячной заработной платы. Так, если в 1991 г. среднемесячная заработная плата работников самой высокооплачиваемой отрасли (газовая промышленность) была в три раза выше, чем у работников с наименьшим уровнем заработной платы (культура и искусство), то в 1-м полугодии 2005 г. разрыв в уровне оплаты труда работников наиболее высокооплачиваемого вида деятельности (добыча топливно-энергетических полезных ископаемых) и низкооплачиваемого (сельское хозяйство, охота и предоставление услуг в этих областях) составил свыше шести раз.

Еще более заметные различия в уровне оплаты труда проявляются при сравнении уровней заработной платы крайних групп работающих в ряду распределения численности работающих по 10%-ным группам. Анализ этой информации показывает, что если в начале 1990-х годов соотношение средней заработной платы работников, попадающих в крайние 10%-ные группы наиболее и наименее оплачиваемых работников, составляло менее восьми раз, в середине 1990-х годов - более 20 раз, то в конце 1990-х годов - превысило 30 раз. С 2001 г. (когда было отмечено наибольшее значение этого показателя) это соотношение снизилось с 40 до 25 раз.

Процессы и тенденции в сфере оплаты труда предопределили необходимость организации новых методов статистического изучения заработной платы. Основные отличия в организации статистики заработной платы в настоящее время от существовавших в дореформенный период принципов состоят в следующем:

1. Введена месячная периодичность сбора сведений об уровне заработной платы и цензовый метод ее представления в зависимости от размера организации;

2. В оперативном режиме осуществляется сбор информации о просроченной задолженности по заработной плате;

3. Введены новые выборочные обследования, программа которых направлена на получение дополнительной информации о заработной плате, спрос на которую формируется под воздействием происходящих изменений в этой области.

Пересмотр системы показателей и методов их получения проводится с учетом приближения их к стандартам, применяемым в международной практике, в соответствии с положениями Конвенции 1985 г. о статистике труда (№ 160) и Рекомендации 1985 г. о статистике труда (№ 170).

В настоящее время информационная база статистики заработной платы формируется с использованием следующих видов статистических наблюдений:

- текущие обследования уровня заработной платы для получения временных рядов в краткосрочном интервале;

- периодически проводимые специальные обследования заработной платы работников и затрат организаций на рабочую силу для получения более подробных сведений о заработной плате работников и других видах затрат работодателей на рабочую силу.

Источником информации о заработной плате наемных работников и затратах организаций на рабочую силу являются сведения государственной статистической отчетности, представляемой предприятиями и организациями.

Целью проведения **текущих статистических наблюдений** является получение данных об уровне и динамике заработной платы в среднем на одного работника в месяц и за отработанный час, по отраслям экономики, видам де-

ятельности, регионам, для исчисления индексов номинальной и реальной заработной платы, а также для международных сопоставлений.

Выборочные обследования заработной платы работников и затрат на рабочую силу проводятся по следующим темам:

- распределение численности работников по размерам среднемесячной заработной платы;
- состав затрат организаций на рабочую силу;
- среднемесячная заработная плата работников органов исполнительной власти;
- среднемесячная заработная плата работников по профессиям и должностям.

По результатам обследования организаций о ***распределении работников по размерам заработной платы*** разрабатывается информация, характеризующая дифференциацию оплаты труда в отраслях экономики (по видам экономической деятельности), субъектах Российской Федерации на основе следующих основных показателей:

- распределение численности работников по размерам заработной платы (по интервальным диапазонам);
- соотношение размеров средней заработной платы работников в крайних 10%-ных группах с наибольшим и наименьшим уровнем заработной платы;
- распределение общей суммы средств, направленных на оплату труда по 10%-ным группам работников;
- численность работников, заработная плата которых ниже величины прожиточного минимума, рассчитанного для трудоспособного населения.

Данное обследование до 1991 г. проводилось один раз в пять лет сплошным методом. С 1994 по 1997 г., в 1999 и 2000 гг. обследование проводилось по выборочному кругу организаций-представителей, характеризующих генеральную совокупность объектов статистического наблюдения (без субъектов малого предпринимательства). Начиная с 2001 г. ежегодно производится научно обоснованная выборка организаций, подлежащих обследованию, данные по которым распространяются на генеральную совокупность объектов статистического наблюдения. Обследуемым периодом является календарный месяц, в последние годы - это апрель.

Статистическое наблюдение за ***составом затрат организаций на рабочую силу*** осуществляется с 1995 г. с периодичностью один раз в два года на выборочной основе во всех регионах Российской Федерации. Статистический показатель затрат на рабочую силу представляет собой издержки, которые несет работодатель в связи с содержанием рабочей силы. Состав затрат на рабочую силу разработан в соответствии с международной классификацией затрат, рекомендованной 11-й Международной конференцией статистиков труда (октябрь 1966 г.) и адаптированной к российским условиям.

В 1999 г. для организации обследования затрат организаций на рабочую силу введена новая модель построения выборочной совокупности с последующим распространением данных, полученных от организаций, на генеральную совокупность организаций в целом по России и регионам по наблюдаемым видам экономической деятельности, формам собственности, группам организаций в зависимости от численности работников.

В бланке обследования затраты организаций на рабочую силу сгруппированы по направлениям затрат в 10 групп, соответствующих Международной стандартной классификации расходов, рекомендованной Международной организацией труда, в частности: оплата за отработанное и неотработанное время; единовременные поощрительные выплаты; оплата питания и проживания; расходы по обеспечению работников жильем; расходы организации на социальную защиту работников; профессиональное обучение; культурно-бытовое обслуживание; налоги и сборы, связанные с использованием рабочей силы.

В составе этих групп выделены показатели, характеризующие структуру заработной платы, натуральную оплату труда, структуру расходов на социальную защиту работников, в том числе взносы в негосударственные пенсионные фонды, на добровольное медицинское страхование.

При анализе затрат организаций на рабочую силу используются следующие основные расчетные показатели: среднемесячные затраты на рабочую силу в расчете на одного работника списочного состава; затраты на рабочую силу в расчете на 1 отработанный час и затраты на 1 оплаченный час.

Выборочное обследование организаций о заработной плате работников по профессиям и должностям

Среди всех видов статистического наблюдения за уровнем заработной платы особое место отводится методам статистического наблюдения за ***средней заработной платой работников по профессиям и должностям***. Для получения информации о заработной плате по отдельным профессиям и должностям Международная организация труда рекомендует странам проводить обследование, которое известно как Октябрьское обследование МОТ. Целью данного обследования является проведение международных сопоставлений покупательной способности заработной платы путем соотнесения данных о среднемесячной и среднечасовой заработной плате работников отдельных профессий с ценами на основные продовольственные товары. С 1985 г. вопросники по Октябрьскому обследованию охватывают 159 профессий из 49 отраслей экономической деятельности.

В России выборочные обследования заработной платы работников отдельных профессий и должностей на систематической основе начали проводиться с 1994 г. Данное обследование проводилось на выборочной основе один раз в два года за октябрь месяц на основе разработанной в соответствии с рекомендациями Международной организации труда и адаптированной к российским условиям методологии.

По результатам обследований данные о месячной и часовой оплате труда, отработанном времени разрабатывались по 187 профессиям и должностям. На основании этих данных проводились расчеты покупательной способности заработной платы работников различных профессиональных групп. Эта информация использовалась для заполнения вопросников МОТ.

Однако в перечне профессий и должностей, который во многом базировался на требованиях вопросника МОТ, далеко не всегда присутствовали те профессии, которые являются типичными для какого-либо отдельного региона и информация о заработной плате которых наиболее важна. Некоторые профессии потеряли свою значимость, в то время как актуальность других возросла. Следует отметить, что вопрос о программе Октябрьского обследования обсуждался на 17-й Международной конференции статистиков труда (ноябрь-декабрь 2003 г.), на которой было принято решение о пересмотре рекомендуемого перечня профессий и должностей в увязке с пересмотром Международного классификатора занятий.

В России пересмотр программы обследования был осуществлен в 2002-2003 гг. При пересмотре программы обследования учитывалась необходимость получения дополнительных данных о размере заработной платы и ее структуре в сочетании с характеристиками работающих (по полу, возрасту, образованию). При разработке программы обследования и методологии его проведения использовался опыт Польши в проведении аналогичного обследования по программе Евростата. В 2003 г. в восьми субъектах Российской Федерации апробирована принципиально новая методология выборочного статистического наблюдения за уровнем заработной платы по профессиям.

В 2004 г. обследование проведено во всех регионах Российской Федерации. Второе полномасштабное обследование проводится за октябрь 2005 г.

Объекты статистического наблюдения. Обследованию подлежат крупные и средние организации со средней численностью работников (включая работающих по совместительству и договорам гражданско-правового характера) 15 человек и выше и работники этих организаций.

Для проведения обследования утверждена форма № 57-т «Сведения о заработной плате работников по профессиям и должностям». В отличие от проводимых ранее Октябрьских обследований, когда предприятие, попавшее в выборку, заполняло сведения по всем работникам на специализированном для данной отрасли бланке с установленным перечнем профессий и должностей, в новом обследовании предусмотрен один бланк для всех видов деятельности с заполнением в свободных строках данных по работникам, попавшим в выборку.

Объектами статистического наблюдения по форме № 57-т являются местные единицы: территориально-обособленные подразделения (включая головные) юридических лиц, а также юридические лица, не имеющие территориально-обособленных подразделений, со следующими видами экономической деятельности: добыча полезных ископаемых; обрабатывающие производства; производство и распределение электроэнергии, газа и воды; строительство; оптовая и розничная торговля, ремонт автотранспортных средств, мотоциклов, бытовых изделий и предметов личного пользования; гостиницы и рестораны; транспорт и связь; операции с недвижимым имуществом, аренда и предоставление услуг (соответственно разделы ОКВЭД: C, D, E, F, G, H, I, K). Эти виды деятельности обследовались в 2004 г. В 2005 г. перечень видов экономической деятельности, подлежащих обследованию, дополнен следующими разделами: M - «Образование», N - «Здравоохранение и предоставление социальных услуг», класс 92 - «Деятельность по организации отдыха и развлечений, культуры и спорта».

Формирование выборочной совокупности. Для формирования выборочной совокупности используется двухступенчатая выборка: на первой ступени формируется выборка организаций, на второй ступени - по отобранной совокупности организаций осуществляется выборка работников.

Генеральная совокупность организаций (основа выборки) для проведения обследования в 2004 г. формировалась на региональном уровне по данным формы № 1-предприятие, в 2005 г. - формы № П-4 за август 2005 г. Переход на использование формы № П-4 для формирования генеральной совокупности осуществлен по предложениям территориальных органов в связи с тем, что форма № 1-предприятие, составленная по итогам за предыдущий год, к моменту составления отчета по форме № 57-т теряет свою актуальность, а также в связи с отсутствием отчетов по форме № 1-предприятие по организациям образования, здравоохранения, культуры и спорта.

Основа выборки (генеральная совокупность) делится на два массива: в первый массив включаются организации с численностью 2000 человек и более, во второй массив - все остальные организации генеральной совокупности. Первый массив обследуется в сплошном режиме. Совокупность организаций второго массива наблюдается в выборочном режиме.

Расслоение (стратификация) второго массива осуществляется по следующим признакам: двум качественным - вид экономической деятельности (39 слоев) и форма собственности (два слоя) и одному количественному - среднесписочная численность работников (не более шести слоев). После расслоения генеральной совокупности по трем признакам осуществляется формирование выборки.

Процедуре отбора организаций предшествуют два этапа, на которых определяется объем выборки организаций, осуществляется его размещение по Нейману в рамках образованных трехмерных слоев и проводится корректировка результатов полученного размещения.

Объем выборки организаций по второму массиву с учетом его размещения в слоях по Нейману рассчитывается по показателю «фонд начисленной заработной платы» с учетом того, что значение коэффициента вариации

оценки (относительной величины стандартной ошибки выборки) по количественному признаку размещения не должно превышать 5%.

Для отбора организаций в образованных трехмерных слоях применяется стандартная процедура случайного отбора.

Общий объем обследуемой совокупности организаций представляет собой сумму объемов двух массивов:

- объем первого массива (обязательно наблюдаемых единиц);

- объем второго массива.

Конечной единицей отбора принимается работник, полностью отработавший обследуемый месяц. Основой выборки на второй ступени является список работников, полностью отработавших обследуемый месяц, который составляют все организации первого массива и организации второго массива, попавшие в выборку. Отбор работников производится из списка, упорядоченного по четырем группам признака «категория персонала»: 1 - руководители, 2 - специалисты, 3 - другие служащие, 4 - рабочие. При этом в начале списка приводится четвертая группа - рабочие, затем первая, вторая и третья группы.

Отнесение работников к категориям персонала производится в соответствии с Общероссийским классификатором профессий рабочих, должностей служащих и тарифных разрядов (ОКПДТР), введенным в действие постановлением Госстандарта России от 26.12.1994 № 367 с 1 января 1996 г. (с учетом последующих дополнений).

Общая численность работников, подлежащих обследованию, устанавливается исходя из численности работников организации, полностью отработавших обследуемый месяц, и находится в интервале от восьми человек (для организаций с численностью до 49 человек) до 64 человек (для организаций с численностью 4000 человек и более).

Для отбора работников применяется систематический отбор, при котором начало отбора определяется случайно. Шаг отбора для первой группы работников в списке (рабочие) определяется исходя из численности работников в списке и числа работников, подлежащих обследованию.

Таким образом, в обследование может попасть работник, имеющий любую профессию (должность).

Программа обследования. Бланк обследования за октябрь 2005 г. состоит из двух разделов. В *первом разделе*, который заполняется на всех работников организации, предусмотрены следующие показатели: среднесписочная численность работников; начислено сумм заработной платы, из них тарифный заработок, выплаты по районному регулированию; количество отработанных за октябрь человеко-часов.

Эти показатели заполняются по категориям персонала (руководители, специалисты, другие служащие и рабочие) отдельно по мужчинам и женщинам.

Во *втором разделе* по каждому работнику, попавшему в выборку, заполняется информация (обезличенная) по следующим показателям: наименование должности, профессии; пол, возраст, образование; общий трудовой стаж; суммы начисленной заработной платы за октябрь всего, в том числе тарифный заработок, выплаты по районному регулированию, другие выплаты; количество отработанных за октябрь часов.

Кодирование наименования *профессий (должностей)* работников, попавших в выборку, и *категорий персонала* производится соответственно пятизначным (код профессии и должности) и однозначным кодом (код категории персонала) с применением Общероссийского классификатора профессий рабочих, должностей служащих и тарифных разрядов.

Формирование итогов обследования осуществляется по *занятиям* в соответствии с Общероссийским классификатором занятий (ОКЗ), принятым постановлением Госстандарта России от 30.12.1993 № 298 и введенным в действие с 1 января 1995 г. Между объектами классификации ОКПДТР и ОКЗ установлены связи, позволяющие относить каждый объект ОКПДТР к соответствующей группировке ОКЗ.

Кодирование *видов экономической деятельности* производится с применением Общероссийского классификатора видов экономической деятельности (ОКВЭД), принятого постановлением Госстандарта России от 6 ноября 2001 г. № 454-СТ.

Образование работника классифицируется по следующим уровням, на основании которых присваивается соответствующий код: высшее профессиональное образование; неполное высшее профессиональное образование; среднее профессиональное образование; начальное профессиональное образование; среднее (полное) общее образование; основное общее; не имеют основного общего.

Распространение результатов обследования на генеральную совокупность и формирование сводных итогов. Распространение выборочных данных обследования на генеральную совокупность производится с использованием общего (агрегированного) веса конечной единицы отбора (работника). Агрегированный вес работника - это величина, обратная значению общей вероятности включения работника в выборку. При двухступенчатом отборе агрегированный вес работника вычисляется по результатам отбора на каждой ступени формирования выборки и равен произведению веса организации (первая ступень отбора) и индивидуального веса работника (вторая ступень отбора).

Вес организации определяется путем деления числа организаций в основе выборки на число отобранных единиц с учетом признаков расслоения (по видам деятельности и формам собственности).

Индивидуальный вес работника определяется делением общей численности работников организации по какой-либо категории персонала, полностью отработавших отчетный месяц, на число работников соответствующей категории персонала, попавших в выборку.

Вес организации используется при распространении данных в целом по организации (1-й раздел Бланка обследования), агрегированный вес - при распространении индивидуальных данных (2-й раздел).

На региональном уровне формируются информационные массивы базы микроданных, содержащие учетные признаки обследованных лиц (без идентификационных признаков) и весовые коэффициенты, рассчитанные для каждой отдельной единицы наблюдения. Эти информационные фонды передаются на федеральный уровень и используются при формировании распространенных данных на федеральном и региональном уровнях.

Программа обследования и применяемая система обработки данных позволяют формировать всестороннюю информацию об уровне заработной платы работников по категориям персонала, группам профессий и должностей работников с учетом их социально-демографических характеристик: полу, возрасту, образованию, что расширяет возможности анализа дифференциации уровня оплаты труда и структуры заработной платы работников по профессиональным группам, видам экономической деятельности, субъектам Российской Федерации.

Расширение программы обследования позволит также обеспечить более полное соответствие рекомендациям Международной организации труда, содержащимся в Конвенции о статистике труда (№ 160), Рекомендации о статистике труда (№ 170) и в соответствующих резолюциях МОТ.

СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ В ОБСЛЕДОВАНИИ ЗАРАБОТНОЙ ПЛАТЫ*

С.В. Степанов, *канд. социол. наук,*
Консалтинговая компания ПЛАНОВА-Консалтинг

При проведении статистических обследований часто возникает ситуация, когда инструментарий обследования (измерения) данных по объективным причинам способен регистрировать данные исследуемого явления только в виде интервалов значений признака, а не его точечных характеристик. Так происходит, к примеру, при измерениях некоторых показателей потока частиц из реактора. Каждый регистратор частиц в силу своих конструктивных особенностей способен регистрировать суммарный импульс, энергию и количество частиц только строго определенного диапазона энергий или скоростей. Необходимое количество регистраторов охватывают весь возможный диапазон и таким образом проводится измерение. Такое свойство некоторых компонент инструментария может объясняться не только объективными обстоятельствами, но и соображениями экономического или организационного характера, в целях сокращения ресурсных затрат при проведении обследования.

Явление регистрируется, и его характеристики измеряются в заранее заданных интервалах. Таким образом, в результате проведенного обследования есть данные по единицам наблюдения, но *отсутствуют данные по аналитическим единицам*, на основании которых можно было бы определить функцию распределения случайной величины, ответственной за вариацию исследуемого явления. В связи с отсутствием представления о *функции распределения в интервалах* затруднены не только аналитические задачи, но также и прогнозные гипотезы о поведении исследуемого явления как за пределами измеренных диапазонов, так и во временной перспективе.

Существенные трудности возникают при определении и расчете статистических показателей для подмножеств аналитических единиц внутри диапазонов измерения и *для подмножеств, охватывающих больше одного интервала*, среди которых есть не целые. Описанные трудности объективной неполноты определения вероятностных характеристик исследуемого явления присущи всем статистическим обследованиям, использующим при получении данных наблюдения идентификацию события, подлежащего регистрации, на основании принадлежности значений некоторого (некоторых) признака (признаков) этого события к заранее заданным интервалам.

Именно такой случай мы встречаем при обработке данных государственного статистического наблюдения по Форме № 1 «Распределение численности работников по размерам начисленной заработной платы», проводимого ежегодно на основании Федеральной программы статистических работ и Производственного плана статистических работ на текущий год. Отчетность предоставляют предприятия, регистрируя данные по численности в заданных интервалах заработных плат и суммы, начисленные работникам, попавшим в соответствующие интервалы.

*Автор выражает благодарность за помощь в подготовке статьи главному специалисту Территориального органа Росстата по Карачаево-Черкесской Республике О.В. Бобрышеву и заместителю отдела Управления статистики труда, образования, науки и культуры О.Б. Жихаревой.

Таблица 1

Фрагмент Формы № 1 «Распределение численности работников по размерам начисленной заработной платы»
Коды по ОКЕИ: человек - 792; рублей - 383;
тыс. рублей - 384

Размер начисленной заработной платы за отчетный месяц, рублей	№ строки	Численность работников - всего, человек	Суммы, начисленные работникам, учтенным в графе 3, рублей
1	2	3	4
До 600,0	01		
600,1-800,0	02		
800,1-1000,0	03		
1000,1-1400,0	04		
1400,1-1800,0	05		
1800,1-2200,0	06		
2200,1-2600,0	07		
2600,1-3000,0	08		
3000,1-3400,0	09		
3400,1-4200,0	10		
4200,1-5000,0	11		
5000,1-5800,0	12		
5800,1-7400,0	13		
7400,1-9000,0	14		
9000,1 -10600,0	15		
10600,1-13800,0	16		
13800,1-17000,0	17		
17000,1-20200,0	18		
20200,1-25000,0	19		
25000,1-35000,0	20		
35000,1-50000,0	21		
50000,1-75000,0	22		
Свыше 75000,0	23		
Всего работников (стр. с 01 по 23)	24		

Данные по единицам наблюдения (предприятиям) агрегируются в виде регламентной отчетности. Аналитической единицей в этом обследовании является работник, однако данных по отдельному работнику в результатах обследования не предусматривает ни методология обследования, ни его инструментарий. Сводные данные представлены в разрезах по формам собственности, отраслям промышленности (виду деятельности) и субъектам административно-территориального деления. Фрагмент сводных данных наблюдения по Карачаево-Черкесской Республике приведен в таблице 2.

Первичные данные, получаемые от предприятий, как это видно из формы отчетности, представляют собой не точечные значения признаков объекта наблюдения, а суммы признака, относящиеся к регистрируемым интервалам, и только к ним. По таким интегрированным данным получить функцию распределения, ее параметры, к примеру, численности для уровней заработных плат, нельзя

Таблица 2

Фрагмент сводной таблицы «Сводные данные о распределении численности, полученные по результатам обследования за апрель 2004 г.» по Карачаево-Черкесской Республике

Размер начисленной заработной платы за отчетный месяц, рублей	№ строки	Численность работников - всего, человек	Суммы, начисленные работникам, учтенным в графе 3, тыс. рублей	Средняя заработная плата, рублей
1	2	3	4	5
До 600,0	01	3160,00	1235714,00	391,10
600,1-800,0	02	3320,00	2348463,00	707,30
800,1-1000,0	03	2999,00	2786625,00	929,20
1000,1-1400,0	04	5006,00	5975929,00	1193,70
1400,1-1800,0	05	4756,00	7560672,00	1589,70
1800,1-2200,0	06	4810,00	9566648,00	1988,80
2200,1-2600,0	07	4627,00	11059946,00	2390,20
2600,1-3000,0	08	4818,00	13388403,00	2778,60
3000,1-3400,0	09	8977,00	28437282,00	3167,80
3400,1-4200,0	10	7548,00	28707707,00	3803,20
4200,1-5000,0	11	5782,00	26738466,00	4624,60
5000,1-5800,0	12	3966,00	21508411,00	5422,60
5800,1-7400,0	13	6908,00	45637542,00	6606,30
7400,1-9000,0	14	2768,00	22332655,00	8068,60
9000,1 -10600,0	15	2044,00	20233797,00	9897,80
10600,1-13800,0	16	985,00	11725323,00	11907,20
13800,1-17000,0	17	240,00	3611545,00	15034,20
17000,1-20200,0	18	170,00	3079813,00	18116,50
20200,1-25000,0	19	102,00	2333304,00	22825,80
25000,1-35000,0	20	29,00	859534,00	29639,10
35000,1-50000,0	21	14,00	576469,00	42701,40
50000,1-75000,0	22	8,00	475301,00	59412,60
Свыше 75000,0	23	20,00	2091658,00	104582,90
Всего работников (стр. с 01 по 23)	24	**73058,00**	**272271206,00**	**3726,80**

без определенных условных допущений. А значит, нельзя достоверно рассчитать статистические показатели по любым разрезам численности.

В составе отчетных таблиц обследования есть «Таблица № 3...», отражающая важную информацию о социально-экономическом состоянии распределения вознаграждения за труд. В этой таблице рассчитываются статистические показатели в иных интервалах, а именно в интервалах десятипроцентных долей (децилей) численности. Пример такой таблицы, точнее ее фрагмент по Карачаево-Черкесской Республике, приведен в таблице 3.

На примере практической расчетной задачи получения вычисляемых данных для таблиц, аналогичных Таблице № 3, требующей преобразования измеренных данных к иным интервалам, рассмотрим необходимое доказательство корректности применяемых расчетных проце-

Таблица 3

Фрагмент Таблицы № 3 «Общие сведения, полученные по результатам обследования за апрель 2004 г.»

Территория 91 Карачаево-Черкесская Республика*											
Все формы собственности В том числе по 10%-ным группам работников											
	Всего	1	2	3	4	5	6	7	8	9	10
Общая сумма средств, направленных на оплату труда, тыс. рублей	272271	4352	8196	12682	16818	21480	24008	28705	35896	46698	73437
Удельный вес средств, направленных на оплату труда, в %	100,00	1,60	3,01	4,66	6,18	7,89	8,82	10,54	13,18	17,15	26,97
Средняя заработная плата, рублей	3726,8	595,7	1121,9	1735,9	2302,0	2940,1	3286,2	3929,0	4913,3	6391,9	10052,0

*Данные, показанные в таблице 3, являются условными, но адекватными реальности структурно.

дур и достаточных допущений, делающих такой расчет возможным. Достоверное преобразование данных, полученных на основе измерений одних интервалов в другие, без знания функции распределения и ее параметров, представляет собой нетривиальную задачу. После завершения статистического наблюдения, сбора и агрегирования данных по всем предприятиям можно оценить распределение численности по всей территории в целом и сделать вывод о форме функции распределения или, как это принято в параметрической традиции, определить принадлежность эмпирического распределения к параметрическому семейству кривых Пирсона. Эту функцию распределения с позиций нашей, интервальной задачи следует назвать *глобальной функцией распределения* (ГФР), чтобы отличить ее от функций распределения в интервалах измерений, которые логично называть локальными ФР. Их взаимосвязь окажется для нас важной с позиций формулирования вычислительной процедуры расчета межинтервальных показателей.

Учитывая интервальный характер исходных данных, для построения графика эмпирического распределения численности по размерам заработной платы воспользуемся в качестве точечной интерпретации интервальных данных показателем средней заработной платы в интервале. Это превращает исходные интервальные измерения в эмпирический ряд из 23 точек.

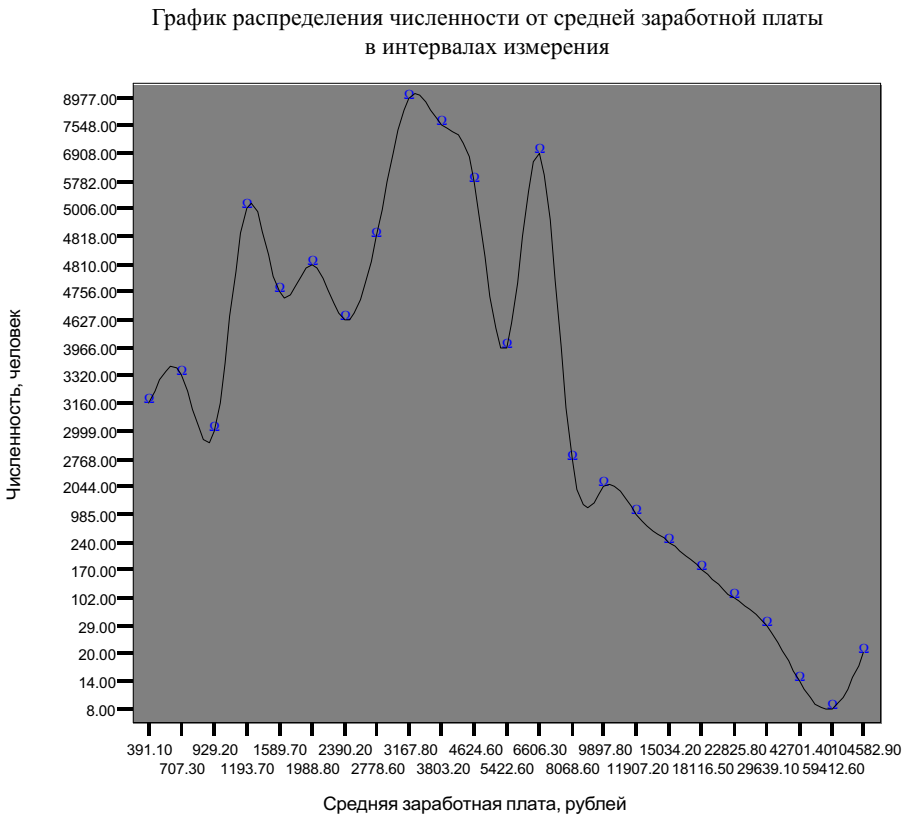

Рис. 1. **Эмпирическое глобальное распределение численности от средней заработной платы**

Эмпирическое распределение, наблюдаемое на рис. 1, можно оценить как логарифмическое нормальное распределение, с учетом того, что по оси X расположены наблюдаемые события (всего числом 23), то есть интервалы наблюдения, а средние зарплаты есть лишь метки этих событий.

Для расчета показателей зарплаты по 10%-ным интервалам численности удобнее представить эту зависимость при ином расположении осей, кроме того, по оси зарплаты отложены не события (измерения), а значения зарплаты.

Распределение средней зарплаты по 10%-ным интервалам накопленной численности

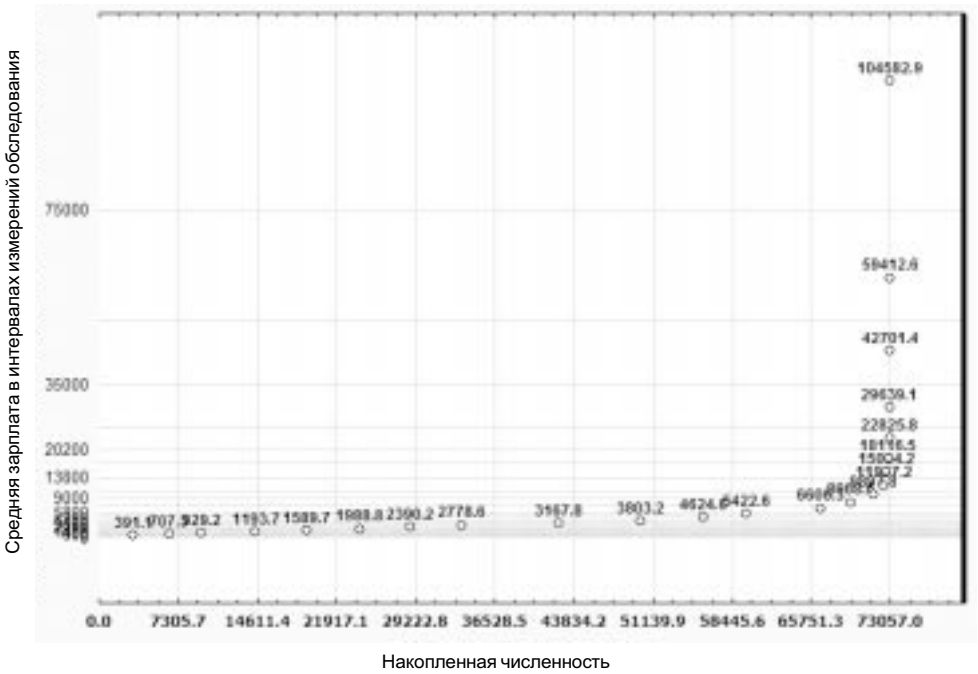

Рис. 2. **Зависимость средней заработной платы от численности**

Средняя заработная плата - вычисляемая величина из начисленного фонда зарплаты и численности в интервалах наблюдения. По оси X можно отложить численность в виде гипотетического ряда отдельных работников, отсортированного по величине их заработной платы. Это соответствует накопленной численности, что не противоречит данным задачи. Исходные данные связаны так:

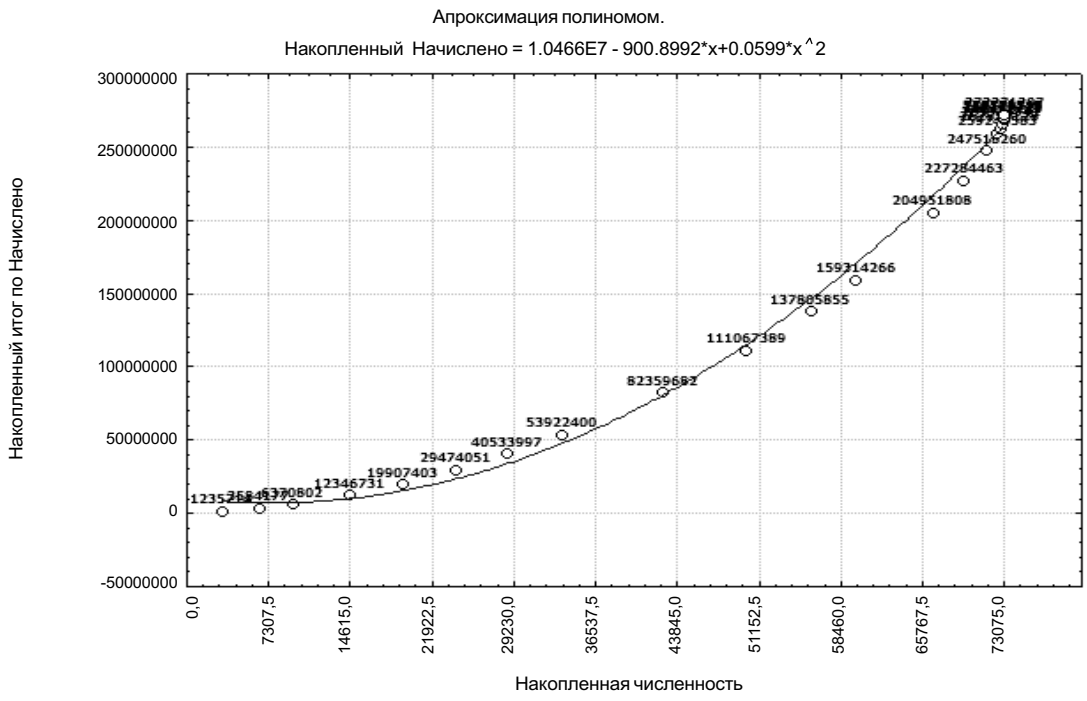

Рис. 3. **Зависимость начисленного фонда заработной платы от численности из наблюдений по 10%-ным интервалам численности**

Или то же в логарифмической шкале с той же численностью и средней заработной платой в интервале измерения. Интервалы измерения и интервалы расчета (10%-ные численности) не совпадают:

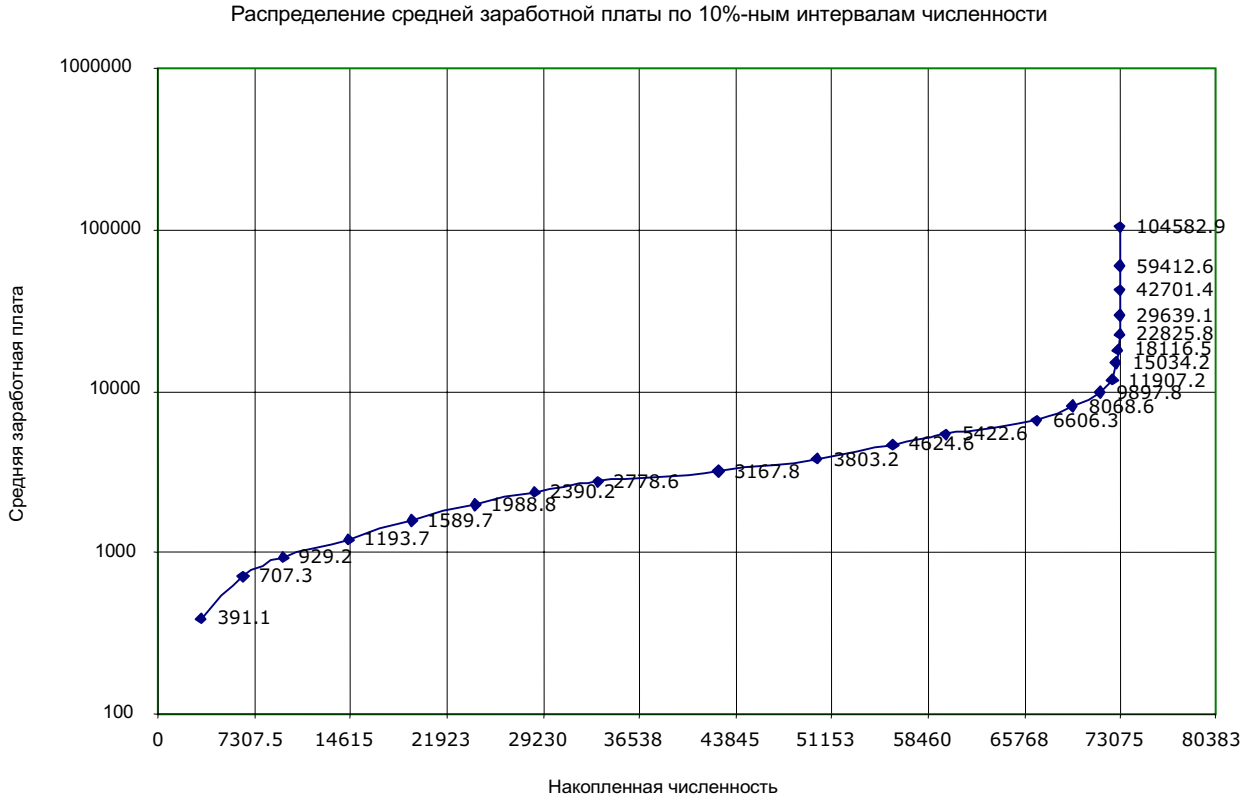

Рис. 4. **Зависимость средней заработной платы от численности в логарифмической шкале**

Конструктивным путем решения задачи создания алгоритма пересчета данных из одних интервалов в другие можно было бы считать получение таких аппроксимирующих функций по ВСЕМ исходным интервалам, *ошибка аппроксимации* по которым *не превосходила бы* или, в лучшем случае, была сравнима с заданной ошибкой выборки. Возможность перерасчета интервалов возможна, если посчитать, что в отсутствии информации о динамике изменения заработной платы, в пределах интервала измерения все заработные платы работников одинаковы и равны средней. Однако такое заведомое упрощение не может считаться нами удовлетворительным приближением реальному положению. Наша задача - попытаться достоверно и доказательно все-таки учесть *неравномерность* распределения в интервале. *Конструктивность* здесь понимается в *интуитивистском* смысле, то есть в возможности получения *эффективного процесса* исчисления интересующего нас показателя, поэтому строго - это требование конструктивности модели представления задано нами ограниченно [5, с. 68; 7, с. 235]. Неконструктивное решение задачи перерасчета внутри и между интервалами измерения возможно, например, в случае использования нейронной сети (НС), обученной на предыдущих наблюдениях, в том числе на иных интервалах. Численные значения внутри интервалов такая обученная НС будет давать, однако такое решение не будет являться в полной мере *конструктивным*, так как у нас не представится возможности записать способ этого решения в каком-либо последовательном формализме для передачи, к примеру, его на реализацию в виде алгоритма, разрешимого по Тьюрингу.

Несмотря на то, что мы не знаем точного вида и параметров зависимости внутри интервалов, мы можем определить несколько ограничивающих правил, которые вытекают из сущности явления и механизма наблюдения. Это создаст условия максимально возможно ограничить класс функций, пригодных для аппроксимаций и достижения *эффективного механизма* расчета показателей фрагмента интервала. Такой механизм позволит производить любые, ограниченные допущениями, перерасчеты внутри и между интервалами. В рассматриваемых нами условиях измерений нельзя даже строго говорить об аппроксимации, так как точечные значения (траектория) внутри интервала нам неизвестны и недоступны. Мы можем сделать лишь некоторые заключения о свойствах класса возможных форм зависимостей, которые позволяет содержание исследуемого явления.

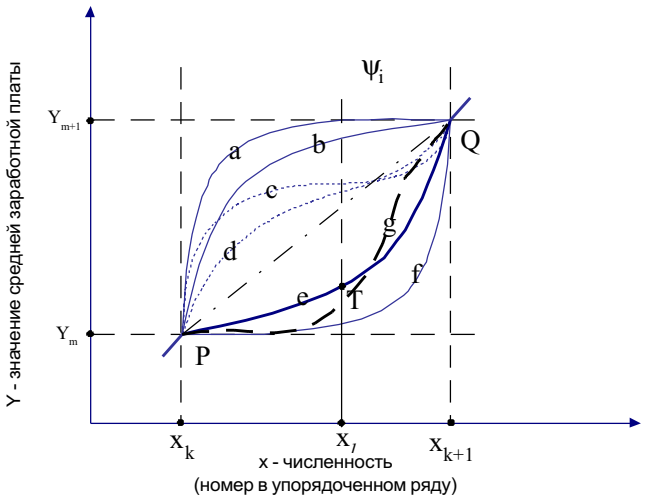

Рис. 5. **Класс функций, допустимых для интерпретации зависимости <<Численность - Заработная плата>>**

На рис. 5 показана обобщенная картина зависимости в пределах одного интервала, где

x_k , x_{k+1} - границы интервала накопленной численности (ряд номеров работников, упорядоченный по размеру заработной платы), в который попали работники, имеющие заработную плату в пределах интервала y_m, y_{m+1};

y_m, y_{m+1} - границы интервала заработной платы, заданные в инструментарии наблюдения, таких интервалов 23;

x_1 - точка внутри рассматриваемого интервала, произвольно его делящая, конкретно - граница 10%-ного интервала численности;

Ψ_i - функции, допустимые для описания зависимости из класса, который *требует определения*.

Исследуемая реальная зависимость может быть представлена линиями функций Ψ_i из некоего обобщенного класса, помеченными латинскими буквами *a, b, c, d, e, f*. Исследуемая нами расчетная модель должна позволять вычислять площади фигур, аналогичных фигурам $\{x_kPTx_1\}$ или $\{x_1TQx_{k+1}\}$.

Запишем формально определение класса функций $\aleph$, такого, что он вполне удовлетворяет описанию исследуемых нами зависимостей, без ограничений тривиальности.

1. По оси X мы отложили накопленную численность, или, что то же самое, перенумеровали работников, расположенных по возрастанию размера заработной платы. Это корректное допущение, несмотря на то, что реально такие данные нам не доступны по условиям проведения наблюдения по Форме № 1. Мы в действительности имеем только интегрированные (суммарные) данные как по численности, так и по накопленной заработной плате в интервале измерения, но при этом не должны забывать, откуда и как эти данные формируются предприятием, составляющим отчет по Форме № 1. Следовательно, функции, определяющие зависимость, могут быть только *неубывающими*, то есть производная такой функции в интервале всегда неотрицательна:

$$\forall x(x \in [x_k, x_{k+1}]) \quad \exists \Psi : \frac{d\Psi}{dx} \geq 0. \qquad (1)$$

2. Интегральные показатели измерений фиксированы по интервалам:

$$\forall i \forall j(i, j \in \Re) \quad \exists \Psi_i \exists \Psi_j : \int_{x_k}^{x_{k+1}} \Psi_i dx = \int_{x_k}^{x_{k+1}} \Psi_j dx = C_k, \qquad (2)$$

где C_k - накопленная заработная плата в интервале (фонд заработной платы).

3. Конструктивное ограничение. Никакая информация о наблюдаемом социально-экономическом феномене не позволяет нам еще более сузить определение класса допустимых функций $\aleph$. По условиям проведения наблюдения такой информации нет. Таким образом, мы попадаем в ситуацию, когда про утверждение, что «такая-то функция, отвечающая требованиям (1) и (2), описывает исследуемую зависимость», мы не можем однозначно сказать, истинно оно или ложно. В самом деле, выбранная нами функция зависимости, удовлетворяющая (1) и (2), не противоречит эмпирическим данным, и, стало быть, утверждение истинно; в то же время мы можем указать сколь угодно много функций, удовлетворяющих тем же условиям, но дающих существенно отличающиеся результаты при расчете внутри интервала площадей фигур вида $\{x_kPTx_1\}$ или $\{x_1TQx_{k+1}\}$ из рис. 5, как это видно для кривых *e* и *g*. Обоснование доказательств в формальных системах без закона «исключенного третьего» не входит в задачу этой статьи, так как это существенно затруднило бы формулирование простого расчетного правила или алгоритма, к которому мы стремимся. Пришлось бы дополнительно обосновывать критерии приемлемости, к примеру энтропийные, или рассчитывать аттрактор траекторий, что поставило бы нас перед требованием больших рядов наблюдений, которое не всегда достижимо. Поэтому мы пойдем на введение *дополнительного ограничения*, имеющего внешний характер в отношении рассматриваемой проблемы и не связанного с логикой, описывающей исследуемое явление, так как мы показали, что она неполна для решения нашей, сугубо расчетной, в конечном итоге, задачи.

Обобщение форм зависимостей, которые мы наблюдаем на рис. 5, позволяет нам выбирать без потери общности между классами гипербол вида:

$$Y = \frac{mx+n}{px+q} \qquad (3)$$

и экспонентами вида:

$$Y = \frac{a}{p} e^{(bx+c)}. \qquad (4)$$

Однако экспоненциальная форма зависимости представляется более предпочтительной. Экспонента - самая

быстрая функция, и, кроме того, в вычислительном смысле упрощается интегрирование, которое нас особенно интересует. В пользу экспоненты, как формы зависимости, наиболее радикально отражающей скорость изменения (дифференциации), можно привести и совсем не математическое обоснование предпочтения. В рассматриваемой задаче речь идет, в частности, об объемах фонда заработной платы, потребляемых группами персонала, существенно различающимися по степени дифференциации заработной платы между ними. Другими словами, одной из целей проводимого наблюдения является определение степени разброса заработной платы между наименее и наиболее оплачиваемыми группами работников. Это важный фактор социального напряжения. Экспоненциальная интерпретация позволяет легко определять по графику зависимости, какова степень и тенденция этого фактора социального напряжения, к какому полюсу ближе ситуация. Чем ближе полученная в наблюдении зависимость к кривой вида *a* из рис. 5, тем, очевидно, дифференциация заработных плат мягче. Заработные платы быстро возрастают с минимума низкооплачиваемого персонала к среднеоплачиваемому и медленно продолжают расти к группе начальников. Напротив, если наблюдаемая зависимость больше похожа на кривую вида f, то можно говорить о возможной тенденции роста социальной напряженности. В этом случае почти весь фонд заработной платы рассматриваемого интервала потребляет высокооплачиваемая группа работников; при этом весьма немногочисленная, а разница в заработных платах от среднеоплачиваемых к начальникам возрастает экспоненциально, то есть очень быстро. Отношение к тому или иному виду экспоненциальной функции определяется параметрами выражения (4), которые вычисляются на основе реальных данных обследования.

Итак, удобство визуальной оценки и простота расчетного правила (алгоритма) дополнительно склоняет нас выбрать для оценки зависимости в интервалах функции класса (4). Интегральное уравнение (3) сокращенно можно конкретизировать так:

$$\int_{x_k}^{x_{k+1}} \frac{a}{p} e^{(bx+c)} dx = C_k . \qquad (5)$$

Решение этого уравнения в конкретных параметрах по всем интервалам x_k и даст нам возможность пересчета на любые диапазоны, не совпадающие с интервалами обследования.

Эта методика оценки параметров распределения корректно обобщается на любые интервалы и позволяет получить если не реальную форму распределения, то как минимум его «динамические» характеристики, что при отсутствии данных внутри интервала существенно увеличивает наше представление о характере исследуемого процесса и нашу возможность оперировать данными субинтервальных диапазонов.

Литература

1. **Вощинин А.П.** Метод оптимизации объектов по интервальным моделям целевой функции. М.: МЭИ, 1987.
2. **Гуссерль Э.** Логические исследования. М.: АСТ, 2000.
3. **Дубров А.М., Лагоша Б.А., Хрусталев Е.Ю.** Моделирование рисковых ситуаций в экономике и бизнесе: Учебное пособие. М.: Финансы и статистика, 1999.
4. **Иванов Ю.Н., Токарев В.В., Уздемир А.П.** Математическое описание элементов экономики. М.: Физматлит, 1994.
5. **Карри Х.** Основания математической логики. М.: Мир, 1969.
6. **Кендалл М., Стюарт А.** Многомерный статистический анализ и временные ряды. М.: Наука, 1976.
7. **Клини С.** Математическая логика. М.: Мир, 1973.
8. **Краснощеков П.С., Петров А.А.** Принципы построения моделей. М.: Фазис, 2000.
9. **Курбатов В.И., Угольницкий Г.А.** Математические методы социальных технологий. М.: Вузовская книга, 1998.
10. **Прохнов Ю.В., Розанов Ю.А.** Теория вероятностей. М.: Наука, 1973.
11. **Пригожин И.** Конец определенности. Москва-Ижевск, R&C Dynamies, 2001.
12. **Шокин Ю.И.** Интервальный анализ. Новосибирск: Наука, 1981.

К ПРОБЛЕМЕ РЕДАКТИРОВАНИЯ ДАННЫХ СЕЛЬСКОХОЗЯЙСТВЕННОЙ ПЕРЕПИСИ

К.Э. Лайкам, *д-р экон. наук,*
заместитель руководителя Росстата,
А.И. Новиков, *канд. экон. наук*

Всероссийская сельскохозяйственная перепись (ВСХП) будет сопряжена со сбором, передачей и автоматизированной обработкой огромных объемов информации - десятков миллионов записей, где каждая запись является вектором $X = (x_1, x_2, ..., x_n)^T$ состояния отдельного хозяйства.

Количество n компонент вектора X зависит от типа хозяйства. Для проведения переписи сформированы 10 списков объектов переписи [1], отнесенных к переписным листам четырех видов:

- «Переписной лист сельскохозяйственных организаций» - форма № 1 (1 СХО);
- «Переписной лист крестьянских (фермерских) хозяйств и индивидуальных предпринимателей» - форма № 2 (2 КФХ);
- «Переписной лист личных подсобных и других индивидуальных хозяйств населения» - форма № 3 (3 ЛПХ);
- «Переписной лист садоводческих, огороднических, животноводческих и дачных некоммерческих объединений граждан» - форма № 4 (4 ОБ).

Первые три формы содержат около 1000 заполняемых

позиций, то есть n = 1000, последняя форма - около 100 позиций (n = 100).

Наибольшими по численности входящих в них хозяйств и некоммерческих объединений являются третья и четвертая категории, которым отвечают переписные листы 3 ЛПХ и 4 ОБ соответственно. В таблице 1 приведены данные о численности хозяйств и объемах произведенной сельскохозяйственной продукции за 2004 г. по первым трем категориям хозяйств.

Таблица 1

Численность хозяйствующих субъектов и объемы произведенной ими продукции по категориям хозяйств

Категория	Число хозяйств	Объем произведенной сельхозпродукции, в %
Сельскохозяйственные предприятия	33000	37,9
Крестьянские (фермерские) хозяйства	264000	4,2
Хозяйства населения	35300000	57,9

Качество статистических данных, которые будут получены в результате ВСХП, определяется большим числом факторов. Важнейшими из них являются достоверность (возможны неумышленные ошибки респондента и сознательное искажение им отдельных показателей) и полнота представления данных (возможны неответы респондента на сложные или «неудобные» вопросы и пропуски отдельных полей в переписном листе переписчиком). Поэтому обработка данных ВСХП будет сопряжена с проблемами обнаружения и исправления ошибок в собранных данных (автокоррекция) и восстановления пропущенных данных. По сложившейся терминологии процедуры автокоррекции несостоятельных данных и импутации будем называть *редактированием данных*.

В странах с большим опытом проведения сельскохозяйственных переписей (Канада, США, некоторые европейские страны) процессам редактирования уделяется повышенное внимание: на редактирование данных может затрачиваться до 40% от общей стоимости переписной кампании. В этих странах разработаны и активно используются при автоматизированной обработке данных сельскохозяйственной переписи пакеты программ: GEIS - статистическое ведомство Канады, AGGIES - департамент сельскохозяйственной статистики США, SLICE - статистическая служба Нидерландов.

Для проведения автокоррекции несостоятельных данных [для обнаружения выбросов (аномальных значений)] в большинстве пакетов реализован метод Хидироглы-Бертелота [2]. Идея метода заключается в следующем. Если значение x_i случайной величины x, имеющей распределение $F(x)$, не принадлежит отрезку $[m - \Delta m_1; m + \Delta m_2]$, то оно является резко выделяющимся и подлежит импутации. Здесь m - медиана случайной величины x; Δm_1, Δm_2 - величины, вычисляемые по значениям квартилей $k_{1/4}$, $k_{1/2}$, $k_{3/4}$ распределения $F(x)$.

Таким образом, если

$$x_i < k_{1/2} - \alpha_1 d\,(k_{1/4}; k_{1/2}) \qquad (1)$$

или

$$x_i > k_{1/2} + \alpha_1 d\,(k_{1/4}; k_{3/4}),$$

где

$$d(k_{1/4};\ k_{1/2}) = \max\{k_{1/2} - k_{1/4}; \beta k_{1/2}\};$$

$$d(k_{1/2}; k_{3/4}) = \max\{k_{3/4} - k_{1/2}; \beta k_{1/2}\},$$

α_1, β - коэффициенты, задаваемые пользователем, то значение x_i идентифицируется как аномальный выброс и подлежит импутации.

В пакете GEIS выделяющиеся значения подразделяются на два класса. В первый включаются аномальные или резко выделяющиеся значения x_i, удовлетворяющие неравенству (1). Их предлагается считать ошибочными и заменять приемлемыми значениями по принятым схемам импутации. Во второй класс включаются значения x_i, удовлетворяющие одному из неравенств:

$$k_{1/2} - \alpha_1 d\ (k_{1/4};\ k_{1/2}) \le x_i < k_{1/2} - \alpha_2 d\,(k_{1/4};\ k_{1/2}); \qquad (2)$$

$$k_{1/2} + \alpha_2 d\ (k_{1/2}; k_{3/4}) < x_i \le k_{1/2} + \alpha_1 d\ (k_{1/2}; k_{3/4})$$

$$\text{и } 0 < \alpha_2 < \alpha_1.$$

Значения x_i, попавшие во второй класс, не аннулируются, но исключаются из списка донорских значений.

Применительно к обработке материалов ВСХП необходимость проверки статистических данных на наличие выбросов не вызывает сомнения. Однако решение о редактировании резко выделяющихся данных, то есть значений x_i случайной величины x, удовлетворяющих одному из неравенств (1), нецелесообразно поручать автоматизированной системе. Система должна выдавать соответствующие предупреждения, а решение о редактировании выбросов должен принимать аналитик.

Подобный вывод следует из результатов обработки материалов пробной сельскохозяйственной переписи по двум регионам - Саратовской области и Краснодарскому краю. В обоих регионах в качестве объекта исследования выступали личные подсобные и другие индивидуальные хозяйства населения. Объемы выборок: 29010 хозяйств в Саратовской области и 16769 хозяйств в Краснодарском крае.

Данные о распределении поголовья крупного рогатого скота, свиней и птицы по двум районам Саратовской области приведены в таблице 2.

Так, в соответствии с неравенствами (1) по показателю «число голов крупного рогатого скота» резко выделяющимися должны быть объявлены 104 значения при $\alpha_1 = 3$ и 23 значения при $\alpha_1 = 5$. Значения квартилей при округлении расчетных значений до целого числа здесь таковы: $k_{1/4} = 1$, $k_{1/2} = 3$, $k_{3/4} = 5$. При $\alpha_1 = 3$ в разряд аномальных, в соответствии с неравенствами (1), попадают значения $x_i > 9$, при $\alpha_1 = 5$ - значения $x_i > 13$.

Таблица 2

Распределение поголовья скота по Саратовской области
(количество хозяйств, имеющих указанное число голов)

Число голов	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	≥20
Крупный рогатый скот	21451	2034	2589	1024	921	390	278	92	81	46	33	20	22	6	9	3	1	1	0	0	9
Свиньи	24180	2529	1154	536	230	177	57	40	40	12	18	12	3	5	2	6	1	1	1	0	6
Птица	18295	1591	3363	1487	1453	598	744	299	357	150	233	81	104	47	57	28	29	10	20	6	50

Нетрудно проверить по данным таблицы 2, что аналогичным является положение и по двум другим показателям. Отметим, что максимальное число голов крупного рогатого скота в одном хозяйстве в Саратовской области равно 51, свиней - 150, птицы - 507. В Краснодарском крае (по двум районам) максимальные значения этих показателей таковы: 20, 53 и 90 голов соответственно.

Таким образом, корректировка, а тем более восстановление резко выделяющихся (экстремальных) значений, являются наиболее сложными вопросами в задаче редактирования. Вероятность принятия тем или иным показателем экстремального значения мала, а вероятность допущения ошибки при его восстановлении, наоборот, велика: относительная ошибка восстановления может составлять сотни процентов. Кроме того, экстремальные значения представляют собой безусловный интерес для содержательного анализа.

Для решения проблемы резко выделяющихся значений необходимо на самом первом этапе - этапе заполнения переписных листов - свести до минимума процент содержания пустых полей в записях с экстремальными значениями. Для этого в ходе ВСХП переписчику будет рекомендовано максимально тщательно заполнять переписной лист, если он обнаружит, что значения определенных показателей превышают заданный пороговый уровень.

Что же касается импутации данных, то известно значительное число методов импутации и способов их классификации. Можно выделить две наиболее часто используемые группы методов: метод донора и метод оценок.

В *методе донора* исходная выборка $\{X_1, X_2,...,X_m\}$ объема m, где $X_i = (x_{1i}, x_{2i}, ..., x_{mi})^T$ - по-прежнему вектор состояния отдельного хозяйства, разбивается по определенным правилам на три класса:

- класс реципиентов;
- класс доноров;
- свободный класс.

В класс доноров включают записи X_i, в которых присутствуют поля x_{ji}, $j = \overline{1, n}$, подлежащие импутации. Донором являются только те векторы X_i, каждое поле x_{ji} которого удовлетворяет всем правилам редактирования. К третьему классу относят записи X_i, которые не содержат полей, подлежащих импутации (не реципиенты), но одновременно не могут быть отнесены к классу доноров, поскольку не проходят контроль на удовлетворение всем правилам редактирования.

В методе донора для каждого вектора X_i из класса реципиентов находится вектор X_k из класса доноров. Он может находиться по методу ближайшего соседа, то есть

$$k_0 = \arg \min_{k \in K} \rho(X_i, X_k),$$

где $\rho(X_i, X_k)$ - расстояние между векторами X_i и X_k в выбранной метрике, или выбираться случайным образом из класса доноров. Импутируемые поля в записи X_i замещаются из записи X_k без каких-либо изменений.

Методы оценок в свою очередь отличаются большим разнообразием. Сюда относят импутацию по предшествующим значениям, импутацию по трендам, парную и множественную регрессию и т. д. Импутируемое значение x_{ji} в методе оценок вычисляется как значение некоторой функции f некоторого числа переменных текущего обследования или многолетних обследований.

Для импутации данных ВСХП предлагается использовать комбинированный метод на основе кластерного анализа, во-первых, для разбиения исходной выборки $\{X_1, X_2,...,X_m\}$ на некоторое число k однородных классов и, во-вторых, для выявления выбросов и решения задачи импутации. Резко выделяющиеся выбросы в соответствии с логикой кластерного анализа должны «проявиться» в отдельных классах экстремальных значений. Задача импутации в этом случае также должна успешно решаться методом ближайшего донора внутри однородного класса.

Эксперименты, проведенные на данных пробной переписи по Саратовской области и Краснодарскому краю, подтвердили справедливость этих предположений. Разбиение исходных совокупностей по каждому региону на классы (кластеры, группы) проводилось методом k-средних [3]. Для этого из множества показателей $\{x_j\}$, $j = 1, n$, входящих в переписной лист 3 ЛПХ ($n \approx 1000$), были выбраны экспертным путем 43 наиболее значимых показателя. По смысловому признаку эти показатели были объединены в следующие группы:

A - трудовые ресурсы (показатели x_1, x_2);

B - земельные ресурсы (показатели $x_3,...,x_6$);

C - площади посевов (показатели $x_7,...,x_{12}$);

D - плодовые деревья, кусты и ягодники (показатели $x_{13},...,x_{18}$);

E - сельскохозяйственные животные (показатели $x_{19},...,x_{26}$);

F - птица (показатели $x_{27},...,x_{32}$);

G - места для содержания скота и птицы (показатели

x_{33},...,x_{37});

H - техника (показатели x_{38},...,x_{43}).

Полный перечень классообразующих показателей с разбиением их на восемь групп приведен в таблице 3.

Таблица 3

Список показателей для кластерного анализа

Условное обозначение группы	Порядковый номер в группе	Обозначение показателя	Наименование показателя
A	1	x_1	Число членов семьи, занятых в хозяйстве, человек
	2	x_2	Численность сезонных работников, человек
B	1	x_3	Общая площадь земли в собственности, га
	2	x_4	Приусадебный земельный участок, га
	3	x_5	Полевые земельные участки, га
	4	x_6	Общая прощадь используемой земли, человек
C	1	x_7	Всего посевов под урожай, кв. м
	2	x_8	в том числе картофель, кв. м
	3	x_9	Овощные и бахчевые культуры открытого грунта, кв. м
	4	x_{10}	Овощи закрытого грунта, кв. м
	5	x_{11}	Кормовые культуры, кв. м
	6	x_{12}	Площадь посевов за пределом участка, кв. м
D	1	x_{13}	Яблони, шт.
	2	x_{14}	Груши, шт.
	3	x_{15}	Слива, шт.
	4	x_{16}	Вишня, шт.
	5	x_{17}	Земляника и клубника, кв. м
	6	x_{18}	Смородина, шт.
E	1	x_{19}	Крупный рогатый скот (КРС), голов
	2	x_{20}	КРС молочного стада, голов
	3	x_{21}	из него коровы, голов
	4	x_{22}	КРС мясного стада, голов
	5	x_{23}	из него коровы, голов
	6	x_{24}	Свиньи, голов
	7	x_{25}	Овцы, голов
	8	x_{26}	Козы, голов
F	1	x_{27}	Птица, голов
	2	x_{28}	Куры яичного направления, голов
	3	x_{29}	Куры мясного направления, голов
	4	x_{30}	Утки, голов
	5	x_{31}	Гуси, голов
	6	x_{32}	Лошади, голов
G	1	x_{33}	Мест для содержания КРС, шт.
	2	x_{34}	Мест для содержания свиней, шт.
	3	x_{35}	Мест для содержания овец и коз, шт.
	4	x_{36}	Мест для содержания лошадей, шт.
	5	x_{37}	Мест для содержания птицы, шт.
H	1	x_{38}	Тракторы, шт.
	2	x_{39}	Плуги тракторные, шт.
	3	x_{40}	Косилки тракторные, шт.
	4	x_{41}	Мотоблоки, мотокультиваторы, шт.
	5	x_{42}	Автомобили грузопассажирские, шт.
	6	x_{43}	Установки доильные, шт.

Разбиение выборочных совокупностей - результатов пробной переписи - по каждому региону производилось с помощью алгоритма *k*-средних на 16, 24 и 32 кластера, реализованного в пакете Statistica 6.0.

Основной вывод по результатам проведенного исследования заключается в том, что получаемые разбиения по обоим регионам оказались близкими по качественному составу. И по Саратовской области, и по Краснодарскому краю четко выделяются кластеры, с одной стороны, с низкими значениями показателей сельскохозяйственного производства, а с другой - с экстремальными значениями отдельных показателей.

Так, по данным пробной переписи, в Саратовской области выделены два кластера с низкими значениями показателей сельскохозяйственного производства. В одном кластере - обозначим его С1 - 11373 хозяйства (39,2% от 29010 хозяйств), в другом кластере С2 - 6658 хозяйств (23%), а всего 18037 хозяйств (62,2%) с очень низкими значениями основных показателей сельскохозяйственного производства (см. таблицу 4).

В среднем хозяйстве первого кластера из 6 соток используемой земли лишь 1,3 сотки отведено под посевы, из них 0,8 сотки заняты картофелем, а 0,5 сотки - овощами и бахчевыми культурами. В хозяйствах данного кластера практически нет скота (227 голов крупного рогатого скота на 11373 хозяйства) и совсем нет техники. Во втором кластере значения одноименных показателей чуть выше, но также мало отличаются от нулевых.

Разбиение выборочной совокупности пробной переписи по Краснодарскому краю на 16 кластеров также содержит кластер С1 с очень низкими значениями показателей по группам E, F, G и H. В нем 9510 хозяйств (56,7% от 16769 хозяйств). В хозяйствах данного кластера также практически нет скота и птицы, нет техники. Но в отличие от аналогичных кластеров С1 и С2 по Саратовской области, значения показателей группы С - посевы под урожай - здесь лишь в 1,5 раза ниже средних по выборочной совокупности. В среднем хозяйстве данного кластера 10,8 сотки земли занято посевами под урожай, из них 9,7 сотки заняты картофелем и 1 сотка - овощами и бахчевыми культурами.

Противоположный полюс в разбиениях выборочных совокупностей по обоим регионам образуют кластеры с экстремальными значениями отдельных показателей и целых групп показателей. К числу таковых по Саратовской области относятся кластеры:

- С11 - максимальное число голов крупного рогатого скота (в среднем хозяйстве кластера 22 головы), овец (46 голов) и птицы (76,5 головы); в составе кластера 14 хозяйств;

- С13 и С16 - максимальные площади земли в собственности (в среднем хозяйстве кластера 47 и 60 га соответственно); в составе каждого кластера четыре и два хозяйства соответственно;

- С15 - кластер в составе трех хозяйств с максималь-

Таблица 4

Средние значения 43 показателей при разбиении на 16 кластеров

Обозначение класса			C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
Число хозяйств			11373	6658	2767	1359	82	1422	723	2599	634	1191	14	38	4	141	3	2
A	1	x_1	1,29	2,69	2,04	1,98	2,43	2,30	2,41	2,47	2,78	3,00	3,43	2,55	2,00	2,50	2,00	4,00
	2	x_2	0,03	0,01	0,02	0,06	0,06	0,02	0,01	0,01	0,01	0,01	0,79	0,05	0,00	0,01	0,00	0,00
B	1	x_3	0,06	0,07	0,15	0,15	0,12	0,09	0,28	0,08	0,12	0,11	0,22	0,17	47,12	0,16	0,09	60,08
	2	x_4	0,06	0,07	0,15	0,15	0,12	0,09	0,27	0,08	0,12	0,11	0,22	0,17	0,12	0,16	0,09	0,15
	3	x_5	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	14,00	0,00	0,00	59,93
	4	x_6	0,06	0,07	0,15	0,15	0,12	0,09	0,28	0,08	0,12	0,11	0,22	0,17	24,86	0,16	0,09	60,08
C	1	x_7	130,91	270,98	1072,28	678,14	627,60	451,24	2386,78	342,57	644,21	466,58	1179,14	1197,55	25416,50	875,96	104,33	1150,00
	2	x_8	82,79	128,65	940,67	408,79	446,89	301,46	2201,13	230,94	442,38	318,46	850,14	974,97	512,50	715,02	30,00	1000,00
	3	x_9	47,61	141,60	126,50	258,80	179,33	148,37	172,65	109,19	197,50	143,97	221,86	186,50	118,00	158,43	74,33	150,00
	4	x_{10}	0,08	0,20	0,01	0,33	0,48	0,13	0,01	0,09	0,12	0,09	0,00	0,00	0,00	0,00	0,00	0,00
	5	x_{11}	0,41	0,52	5,09	10,21	0,90	1,28	13,00	2,36	4,21	4,06	107,14	36,08	36,25	2,51	0,00	0,00
	6	x_{12}	0,12	0,16	1,16	0,29	0,00	1,14	1,38	0,62	0,00	0,00	0,00	0,16	0,00	0,00	0,00	0,00
D	1	x_{13}	0,54	1,12	1,11	3,80	3,99	1,35	1,34	0,86	2,11	1,43	1,00	2,32	4,00	1,70	2,67	4,50
	2	x_{14}	0,03	0,10	0,07	1,06	0,45	0,14	0,16	0,07	0,28	0,13	0,14	0,26	0,50	0,16	0,33	0,00
	3	x_{15}	0,19	0,53	0,50	2,56	1,45	0,59	0,64	0,48	1,29	0,63	0,07	1,08	1,50	0,79	1,00	0,50
	4	x_{16}	0,43	0,95	0,87	5,30	1,88	1,20	0,94	0,95	2,62	1,47	0,86	1,42	2,00	1,72	0,00	1,00
	5	x_{17}	0,89	3,39	2,13	19,17	13,00	3,60	1,95	2,52	6,90	4,04	1,29	5,45	1,50	4,26	10,67	0,00
	6	x_{18}	0,95	2,75	1,98	25,12	8,28	2,91	2,43	2,46	5,85	4,01	0,00	3,13	12,50	4,02	0,00	0,50
E	1	x_{19}	0,02	0,04	0,69	0,64	0,96	1,06	1,88	2,05	2,88	5,46	22,07	3,05	1,50	2,53	0,00	7,00
	2	x_{20}	0,02	0,04	0,69	0,63	0,96	1,05	1,88	2,05	2,84	5,43	21,64	3,05	1,50	2,49	0,00	7,00
	3	x_{21}	0,01	0,00	0,36	0,30	0,48	0,45	0,90	1,12	1,15	2,30	6,29	1,32	0,75	1,08	0,00	3,50
	4	x_{22}	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,02	0,03	0,43	0,00	0,00	0,04	0,00	0,00
	5	x_{23}	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
	6	x_{24}	0,06	0,16	0,60	0,72	0,73	1,36	1,24	1,34	3,27	4,05	2,14	2,61	3,00	2,33	0,00	1,00
	7	x_{25}	0,02	0,04	0,17	0,38	0,06	0,51	0,1148	0,55	1,41	2,82	46,00	1,45	0,00	1,11	0,00	0,00
	8	x_{26}	0,02	0,04	0,07	0,04	0,06	0,06	0,16	0,02	0,10	0,21	0,71	0,00	0,00	0,14	0,00	0,00
F	1	x_{27}	1,06	3,91	6,38	8,91	19,28	31,29	14,38	7,75	49,18	20,88	76,50	21,71	10,00	18,67	369,00	33,50
	2	x_{28}	0,71	2,54	4,13	5,82	9,29	27,13	9,86	4,84	18,95	12,98	35,71	9,82	3,75	10,82	33,33	0,00
	3	x_{29}	0,21	0,79	1,16	1,69	5,28	1,88	1,79	1,37	4,32	3,00	7,50	7,26	1,00	2,93	233,33	20,00
	4	x_{30}	0,10	0,41	0,68	0,90	3,21	1,09	1,59	1,02	23,45	2,73	22,36	3,29	3,75	2,70	2,33	7,00
	5	x_{31}	0,04	0,15	0,36	0,42	1,50	1,09	0,89	0,52	2,19	1,93	10,21	0,29	1,50	2,16	100,00	6,50
	6	x_{32}	0,00	0,00	0,04	0,03	0,01	0,04	0,12	0,08	0,12	0,28	2,71	0,13	0,00	0,11	0,00	0,00
G	1	x_{33}	0,19	0,20	1,34	1,26	1,26	1,68	2,63	2,68	3,53	5,66	18,79	4,00	2,25	3,33	0,00	7,00
	2	x_{34}	0,32	0,60	1,59	1,89	1,67	2,55	2,31	2,68	5,02	5,32	2,86	5,03	4,75	3,81	0,00	1,00
	3	x_{35}	0,14	0,18	0,48	0,94	0,66	0,93	1,07	1,10	1,94	3,46	46,14	1,39	0,00	1,93	0,00	0,00
	4	x_{36}	0,01	0,01	0,06	0,04	0,01	0,05	0,16	0,09	0,13	0,28	3,00	0,13	0,00	0,13	0,00	0,00
	5	x_{37}	4,20	9,17	15,77	16,10	21,98	39,02	24,40	14,90	49,25	24,59	59,86	30,24	18,75	26,33	440,00	33,50
H	1	x_{38}	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,07	0,82	0,00	1,01	0,00	0,50
	2	x_{39}	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,00	0,14	0,00	0,00
	3	x_{40}	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,50
	4	x_{41}	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,11	0,00	0,00	0,00	0,00
	5	x_{42}	0,00	0,02	0,01	0,02	0,05	0,02	0,04	0,01	0,03	0,02	0,21	0,11	0,00	0,07	0,33	0,00
	6	x_{43}	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,02	0,01	0,00	0,03	0,00	0,00	0,00	0,00
Имя кластера			OOA	OOC	CBF	DCB	HCD	OCF	CBE	OCG	FDC	EHF	EGF	HDC	DB	HEC	FGH	ADE

ным числом голов птицы (в двух хозяйствах по 300 голов, а в одном - 507 голов).

В Краснодарском крае выделены следующие кластеры с экстремальными значениями показателей:

- С12 - максимальное число голов крупного рогатого скота - в среднем хозяйстве кластера 8,4 головы; в состав кластера входят 88 хозяйств;

- С13 - максимальные площади земли в собственности (276,4 га в среднем хозяйстве кластера); максимальные площади посевов под урожай (9,2 га); в составе кластера четыре хозяйства;

- С14 и С15 - максимальные площади земли под кормовыми культурами (4,5 га) и под картофелем (1,2 га) соответственно; в кластере С14 - 12 хозяйств, в С15 - 18 хозяйств;

- С16 - максимальное число голов крупного рогатого скота мясного направления (6 голов) и максимальное число голов птицы (90 голов); в составе кластера одно хозяйство.

По данным пробной переписи, хозяйства с очень низкими значениями основных показателей сельскохозяйственного производства в Саратовской области составляют 62,2% от всей выборочной совокупности и 56,7% - в Краснодарском крае. Число хозяйств с экстремальными значениями показателей в обоих регионах не превышает 1%. Оставшиеся 37% ...43% индивидуальных хозяйств населения образуют массив «средних» хозяйств. Он, естественно, неоднороден и в результате применения кластерного анализа разбивается на относительно однородные группы со своей спецификой. Так, например, в Саратовской области и в Краснодарском крае выделены кластеры в количестве 82 и 197 хозяйств соответственно. Значения основных показателей в них близки к средним по всей выборочной совокупности, но объединяет их то, что каждое хозяйство кластера имеет мотокультиватор или мотоблок.

Все приведенные выше результаты соответствуют разбиениям выборочных совокупностей на 16 кластеров. При увеличении числа кластеров до 24 происходило деление некоторых кластеров, полученных при разбиении совокупности на 16 кластеров, на более мелкие, с более резким «подчеркиванием» особенностей кластеров, и прежде всего кластеров с экстремальными значениями показателей.

Были проведены эксперименты по восстановлению количественных данных методом ближайшего соседа с использованием результатов кластерного анализа по следующей схеме. Для каждого вектора X_i из базы реципиентов сначала находился кластер C_{k0}, ближайший к вектору X_i в соответствии с правилом

$$k_0 = \arg \min_{1 \le k \le 16} \rho(X_i, \overline{X}_k),$$

где X_k - вектор средних значений 43 показателей в k-м кластере. Пустые поля x_{ji} вектора X_i заполнялись при этом для каждого k соответствующими средними значениями x_{ik} из k-го кластера. Затем внутри кластера C_{k_0} находился ближайший вектор X_{k_s}, соответствующие поля которого импутировались в запись X_i.

Наилучшие результаты восстановления по этой схеме были получены в кластерах с низкими и средними значениями всех компонент вектора-реципиента. Неудовлетворительные результаты восстановления, и это достаточно естественно, оказались в кластерах с экстремальными значениями показателей. В таких кластерах мал объем донорской базы и часто высока вариация показателей, не участвовавших в формировании кластера.

Для увеличения точности процедур импутации предполагается, наряду с описанной схемой восстановления значений количественных показателей, использовать функциональные и корреляционные взаимосвязи между отдельными показателями.

Импутация неколичественных признаков будет проводиться по индивидуальным для каждого такого признака правилам редактирования. Эти правила строятся по многоступенчатой логической схеме с основными блоками вида: «если ..., то ....».

В целом результаты проведенных на материалах пробной сельскохозяйственной переписи исследований показали возможность успешного решения задач автокоррекции и импутации на основе представленных подходов и алгоритмов для подавляющего числа личных подсобных хозяйств страны.

Литература

1. **Субботина Л.В.** Об основных методологических положениях проведения пробной сельскохозяйственной переписи 2004 года в Российской Федерации // Вопросы статистики. 2005. № 7.

2. **Hidiroglou M.** and **Berthelot J.** Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, № 12, 1986. P. 73-83.

3. **Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.** Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.