УТВЕРЖДЕНЫ
приказом Росстата
от 19.04.2013 № 165

Методологические положения

по формированию массивов деперсонифицированных микроданных годового структурного обследования по форме федерального статистического наблюдения № 1-предприятие «Основные сведения о деятельности организации» общего пользования для представления пользователям в аналитических целях

Содержание

Введение.

Основной целью государственной статистики является обеспечение информационных потребностей государства и общества в полной, достоверной, научно обоснованной и своевременно предоставляемой официальной статистической информации. Согласно статье 2 Федерального закона от 29.11.2007 № 282-ФЗ (ред. от 16.10.2012) «Об официальном статистическом учете и системе государственной статистики в Российской Федерации» официальная статистическая информация представляет собой сводную агрегированную документированную информацию о количественной стороне социальных, экономических, демографических, экологических и других общественных процессов в Российской Федерации, формируемую субъектами официального статистического учета в соответствии с официальной статистической методологией. Агрегирование данных включает формирование общих итогов по всей совокупности наблюдаемых единиц, а также ее разграничение на группы в соответствии с действующими общероссийскими классификациями.

При этом субъекты официального статистического учета гарантируют респондентам конфиденциальность полученных от них индивидуальных данных по показателям, содержащимся в формах государственного статистического наблюдения, и используют эти данные только для формирования официальной статистической информации (ст. 9 федерального закона № 282-ФЗ), что соответствует основополагающим принципам официальной статистики, принятыми Статистической комиссией ООН в 1994 г.

Статистика, как отрасль знаний, предоставляет инструмент, позволяющий выявлять и измерять закономерности развития социально-экономических явлений и процессов, взаимосвязи между ними. Это очень важно при проведении научных и аналитических исследований, построении экономических моделей, принятии управленческих решений. Однако, агрегированных данных, предоставляемых в виде официальной статистической информации, бывает недостаточно для выявления множества однокачественных варьирующих явлений.

Федеральная служба государственной статистики, в соответствии с возложенными на нее полномочиями, представляет в установленном порядке официальную статистическую информацию органам государственной власти и местного самоуправления, средствам массовой информации, научным и другим организациям и гражданам.

Международная статистическая общественность обратила внимание на потребность в получении и возможности предоставления обезличенных статистических данных респондентов. Это значительно расширяет перечень пользователей информацией, подтверждает качество официальной статистической информации и улучшает имидж государственной статистики при сохранении доверия респондентов. Многие национальные статистические

службы (Австралии, Финляндии, Нидерландов, Швеции, США и др.) представляют сообществу исследователей набор обезличенных персональных данных. Представление данной информации нормативно закреплено, определены регламенты доступа пользователей к информации, разработаны соответствующие правила ее обезличивания.

Настоящие методологические положения разработаны с учетом международных рекомендаций в области распространения микроданных внешним пользователям и научных исследований с целью расширения возможностей использования статистических микроданных годового структурного обследования широким кругом исследователей в аналитических целях при обеспечении конфиденциальности данных.

1. Основные понятия и определения

Микроданные - набор единичных записей об индивидуальном объекте (респонденте), каждая из которых содержит набор переменных (показателей) в отношении данного объекта. Четыре категории переменных (необязательно являются непересекающимися):

- прямые идентификаторы,
- косвенные идентификаторы,
- конфиденциальные переменные,
- неконфиденциальные переменные.

Деперсонификация микроданных (анонимизация микроданных) – процедура защиты (маскировки) конфиденциальных данных от раскрытия с применением определенных методов.

Ре-идентификация –

происходит, когда на основе сравнения значений идентифицирующих переменных единица i‘ из внешнего файла определена как соответствующая единице i в массиве микроданных, и установлено, что данная связь является корректной.

Категориальные переменные (данные) - переменные, принимающие значения из некоторого ограниченного набора категорий, связанных с неисчисляемыми признаками, такими как названия (товаров, услуг и др.), выходные переменные в классификационных моделях (метки классов).

Количественные (численные) переменные (данные) - переменные, которые регистрируются с помощью чисел, имеющих содержательный смысл. С количественными переменными можно выполнять все обычные операции над числами, такие, как вычисление среднего и др.

Выделяют два типа количественных переменных: дискретные и непрерывные.

Дискретная - это переменная, которая может принимать значения только строго определенные значения из некоторого списка определенных значений, например, целочисленные.

В отличие от дискретных переменных непрерывные переменные могут принимать любое значение в пределах определенного числового интервала. Исчисления производятся только с непрерывными переменными.

Прямые идентификаторы - переменные, которые однозначно идентифицируют респондента. Например: регистрационный код организации, ее наименование, адрес и т.п.

Косвенные идентификаторы (ключевые переменные) – переменные, которые идентифицируют респондента с той или иной степенью неопределенности. Тем не менее, комбинация косвенных идентификаторов может дать однозначную идентификацию. Например: вид экономической деятельности, населенный пункт, численность работников.

Конфиденциальные переменные – переменные, которые содержат деликатную информацию о респонденте. Например: объем производства, финансовые показатели деятельности организации.

Неконфиденциальные переменные – переменные, которые не относятся ни к одной из вышеперечисленных категорий.

Модификация данных - искажение массива микроданных перед тем, как предоставить к нему доступ.

Сокращение данных - частичная фильтрация (удаление) данных или снижение уровня детализации исходного массива данных.

Абсолютно анонимные микроданные – статистические данные, обработанные методами контроля раскрытия статистической информации путем удаления отдельных переменных и модификации данных до такой степени, что идентификация респондентов является невозможной.

Де-факто анонимные микроданные.

Микроданные являются де-факто анонимными, если нельзя полностью исключить раскрытие конфиденциальных данных, но это может произойти только вследствие чрезмерно затраченного времени, вложения значительных средств и людских ресурсов. Де-факто анонимность микроданных зависит не только объема сохранившейся в данных информации, но и от возможностей, существующих для идентификации объекта статистического наблюдения. Решающее значение имеет наличие дополнительных знаний об индивидуальном объекте и то, каким образом эти данные используются.

Формально обезличенные микроданные - удаление прямых идентификаторов объекта, при этом косвенные идентификаторы (например, виды экономической деятельности, территориальная принадлежность), а также наблюдаемые переменные в основном сохраняются.

Риск и полезность

Методы и решения в области контроля раскрытия статистической информации для минимизации риска раскрытия должны обеспечивать максимальную полезность статистических данных. Задача заключается в том, чтобы найти разумный баланс: сохранить полезность информации и при этом

обеспечить, чтобы риск раскрытия не превышал максимально допустимого уровня.

Годовое структурное обследование – федеральное статистическое наблюдение по форме № 1-предприятие «Основные сведения о деятельности организации». Проводится Федеральной службой государственной статистики ежегодно. Обследованию подлежат юридические лица всех форм собственности, являющиеся коммерческими организациями, а также некоммерческие организации, осуществляющие производство товаров и услуг для продажи на сторону, (кроме субъектов малого предпринимательства, бюджетных организаций, банков, страховых и прочих финансовых и кредитных организаций).

2. Методы деперсонификации микроданных и основные критерии оценки их эффективности

Методы деперсонификации решают задачу защиты микроданных, которая заключается в предотвращении привязки конфиденциальной информации к конкретной единице наблюдения. С их помощью защищенный массив микроданных можно получить путем маскировки исходных данных, то есть, сгенерировав модифицированную версию исходного массива микроданных.

Методы разделяют на два типа:

- Непертурбативные (сокращение данных): методы не предусматривают модификации данных, но выполняют частичную фильтрацию (удаление) данных или снижение уровня детализации исходного массива данных.
- Пертурбативные (модификация данных): массив микроданных искажается перед тем, как предоставить к нему доступ. Использовать пертурбационные методы следует таким образом, чтобы статистические характеристики, рассчитанные на базе модифицированного массива, не слишком отличались от рассчитанного из оригинального массива данных.

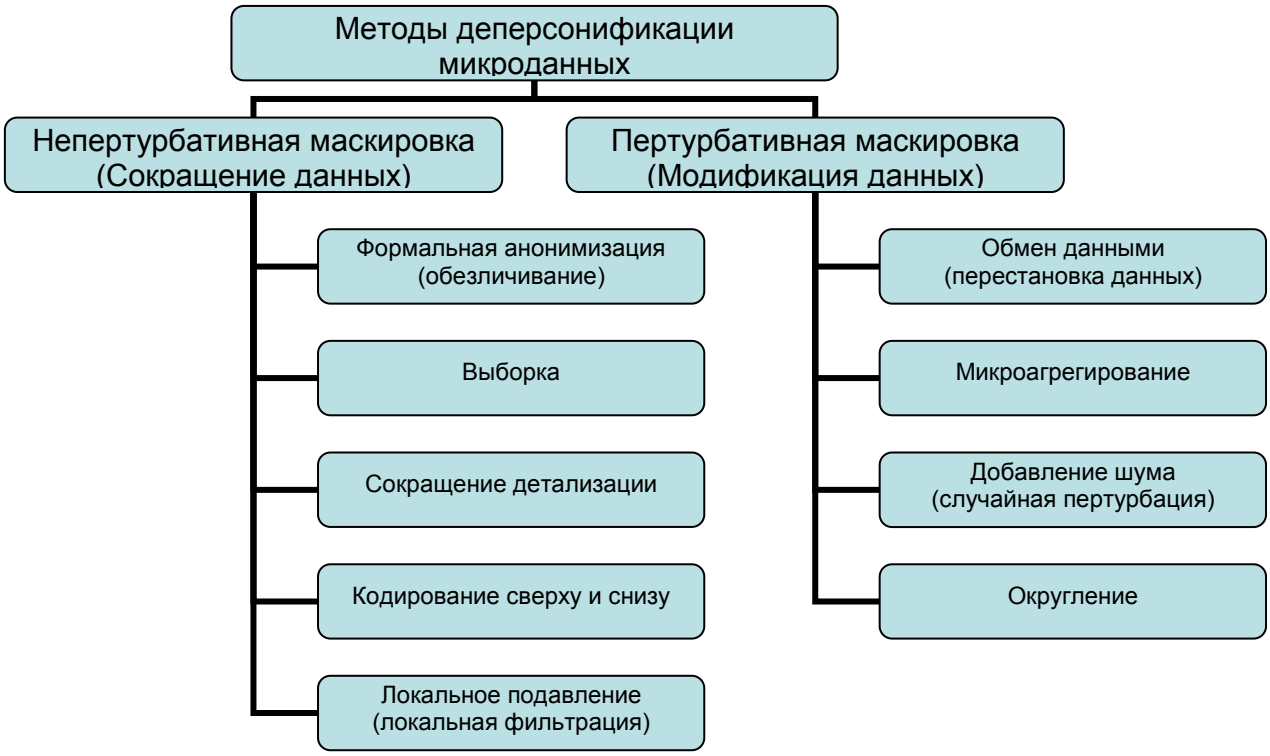

Рис. 1 - Методы деперсонификации данных

3. Общее описание алгоритмов деперсонификации при применении различных ее методов.

Формальная анонимизация (обезличивание)

Формальная анонимизация (обезличивание) заключается в удалении из каждого вектора данных формальных или прямых идентификаторов объекта. После обезличивания объект может быть однозначно опознан только по косвенным идентификаторам.

Формальная анонимизация микроданных является обязательной процедурой при предоставлении доступа к микроданным.

Выборка

В случае использования выборки публикуется не исходный файл микроданных, а выборка S из оригинального массива данных. Данный метод предусматривает публикацию микроданных только для случайной выборки данных. Метод требует последующего применения пертурбационных методов.

Сокращение детализации

Метод заключается в снижении информативности микроданных путем сокращения их детализации, которое может быть достигнуто путем увеличения масштаба шкалы измерения переменной или сокращением числа категорий, которыми представлен каждый косвенный идентификатор.

Для категориальной переменной V_i объединяем несколько категорий с целью формирования новых (менее конкретных) категорий; в результате получаем новую переменную V'_i для которой $|D(V'_i)| < |D(V_i)|$, где $|\ \ |$ – обозначение мощности множества. Для численной переменной глобальное перекодирование означает замену V_i на новую дискретную переменную V'_i. Иными словами, потенциально бесконечный интервал $D(V_i)$ отображается на конечном интервале $D(V'_i)$.

Применительно к численным переменным сокращение детализации может быть выполнено путем замены метрической шкалы на ординальную или интервальную шкалу или уменьшением детализации переменных.

Кодирование сверху и снизу

Кодирование сверху и снизу – особый случай перекодирования; эту методику можно использовать для ранжируемых переменных – т.е. для численных и для категориальных ординальных переменных. Суть заключается в том, что верхние значения (превышающие некий порог) группируются для формирования новой категории. То же самое проделывается с нижними значениями (не превышающими некоего порога).

Локальное подавление

Метод локального подавления (локальной фильтрации) используется для микроданных в случаях, когда экстремальное значение (выделяющееся наблюдение) переменной или экстремальная комбинация значений переменных присутствуют в одном или более векторах данных. Экстремальное значение или экстремальная комбинация значений подавляется, так как их наличие значительно упрощает процедуру идентификации объекта, особенно в тех случаях, когда экстремальными являются значения косвенных идентификаторов. Используются два варианта метода подавления:

- пропуск всех экстремальных значений или комбинаций значений, которые присутствуют в индивидуальных данных, и замена их на “пропущенное” значение. При этом пользователь статистики будет знать, что пропущено экстремальное значение, но не будет владеть реальной цифрой, так как не известна степень и направление “экстремальности”, т.е. велико или мало экстремальное значение, и насколько оно велико или мало;

- удаление всего вектора данных. Этот вариант используется в случае, когда данные содержат очень необычное значение или комбинацию значений, особенно в случаях, когда данный объект широко известен.

Оба варианта метода локального подавления производят отклонение данных, так как оценки величины, полученной на основе микроданных, в которых некоторые значения были подавлены, будет отличаться от оценки, рассчитанной на основе реальных данных.

Подавление, как и другие методы, основанные на сокращении данных, снижает качество данных для проведения анализа.

Обмен данными

Метод обмена данными (свопинг данных или многомерная трансформация) основан на модификации данных. Суть подхода заключается в том, чтобы преобразовать базу данных, поменяв местами значения конфиденциальных переменных индивидуальных записей. Эта перестановка осуществляется таким образом, чтобы частоты низкого порядка сохранялись в том же состоянии.

Другой вариант перестановки данных в массивах микроданных – перестановка рангов. Значения переменной X_t ранжируются в порядке возрастания, затем каждое ранжированное значение X_t меняется местами с другим значением, случайно выбранным в некотором ограниченном диапазоне.

С целью избежать избыточной защиты (модификации) данных для модификации только изолированных единиц наблюдения используется метод вменения значений ближайшей кластерной единицы. Вменение следует применять с использованием значений ближайших (относительно функции

расстояния, использованной в алгоритме кластеризации) не подверженных риску соседей, в противном случае увеличение неопределенности может оказаться недостаточным.

Метод сводится к следующей процедуре:

1. Пусть $x^{isolated}$ – значение, которое требуется защитить.

2. Находим ближайшую кластерную единицу x_p , для которой:

$d(x^{isolated}, x_p) = \min_{x_c \in C}\left(d\left(x^{isolated}, x_c\right)\right)$, где C – множество всех кластерных единиц.

3. Защищенное значение $x'^{isolated}$ принимает значение x_p.

Микроагрегирование

В основе микроагрегирования лежит положение о том, что существующие правила в отношении конфиденциальности разрешают публиковать агрегированные данные, если записи соответствуют группам в составе k или более объектов (принцип k-анонимности), ни один из этих объектов не является доминирующим в группе (т.е. не определяет групповые показатели), а k – пороговое значение. Строгое соблюдение таких правил конфиденциальности обусловливает выполнение подмены индивидуальных значений значениями, рассчитанными для малых множеств (микроагрегатов). Это базовый принцип микроагрегирования.

Для получения микроагрегатов исходная совокупность единиц наблюдения определенным образом разделяется на небольшие группы ближайших друг к другу объектов размером не менее k. *Классические алгоритмы микроагрегирования* требуют, чтобы все группы (возможно, кроме одной) имели размер k. Если количество всех объектов N кратно k, то создается $n=N/k$ групп по k объектов в каждой. Если N не кратно k, то последняя группа, содержащая менее k объектов, объединяется с предыдущей и, таким образом, содержит более, чем k объектов. Затем для каждой группы рассчитывается среднее значение переменной, после чего это значение используется вместо оригинальных данных для всех единиц данной группы. Таким образом, реальный объект заменяется некоторым суррогатным объектом. Особое внимание при этом должно уделяться выделяющимся наблюдениям (объектам), по тем или иным показателям значительно отличающимся от других.

Добавление шума

Добавление шума или метод случайной пертурбации данных хорошо подходит к численным переменным. Основная идея метода заключается в том, что к истинному значению переменной x_i добавляется «шум» в виде случайной величины e, и затем истинное значение заменяется модифицированном значением: $x_i' = x_i + e$.

Распределение случайной величины *e* выбирается отдельно для каждого конкретного случая. В общем случае величина *e* должна иметь нулевое среднее: *E(e) = 1*. При этом отрицательные модифицированные значения показателя, получаемые при добавлении отрицательного «шума», заменяются на 0 при условии, что отрицательные значения переменной недопустимы.

Уровень защиты данных определяется величиной вводимого «шума». Чем больше дисперсия *D(e)* случайной величины *e*, тем сильнее модифицируются исходные данные. Это увеличивает надежность защиты данных от раскрытия, но тем сильнее искажения данных в результате пертурбации.

Округление

Данная методика основана на замене исходных значений переменных округленными значениями. Для переменной X_t, округленные значения выбираются из массива точек округления, которые определяют массив округления. В исходном многомерном массиве данных округление переменных обычно выполняется поочередно (одномерное округление), хотя возможно и многомерное округление. Базовый принцип округления позволяет использовать эту методику для численных данных.

При формировании массивов деперсонифицированных годового структурного обследования по форме № 1-предприятие обработка данных производится в два этапа – после использования методов, основанных на сокращении, применяются методы модификации данных. При этом в качестве методов, основанных на сокращении, используется выборка, локальное подавление, сокращение детализации, а в качестве методов модификации – микроагрегирование, обмен данными и случайная пертурбация (добавление шума). На конечном этапе может применяться округление.

4. Правила обеспечения прямой и косвенной конфиденциальности данных годового структурного обследования

4.1. Риски раскрытия информации

Риск идентификации определяется как вероятность того, что пользователь идентифицирует хотя бы один объект в массиве микроданных, что может привести к раскрытию конфиденциальной информации о данном объекте. Возможны следующие способы идентификации хозяйствующих субъектов:

- идентификация через сопоставление записей об индивидуальных объектах с использованием внешней информации, к которой пользователь имеет доступ (сценарий «Внешний реестр»);

■ непреднамеренная идентификация единицы наблюдения (сценарий «Спонтанная идентификация»);

■ возможность частичного раскрытия данных первично защищенных объектов путем привлечения и сопоставления агрегированных данных, приведенных по различным группировкам, публикуемым в открытых источниках агрегированных статистических данных (сценарий «Вторичная идентификация»).

4.2. Правила обеспечения конфиденциальности информации

Файл микроданных публикуется не ранее чем через два года после отчетного периода.

Правила обеспечения прямой конфиденциальности.

1. Прямые идентификаторы «Наименование отчитывающейся организации», «Код отчитывающейся организации (по ОКПО)», «Код вышестоящей организации (по ОКПО)» и «Почтовый адрес» удаляются из файла микроданных.

2. Переменные, которые могут привести к спонтанной идентификации предприятий, а также идентификации на основании сведений из внешних источников, удаляются из файла микроданных. Их перечень применительно к годовому структурному обследованию приведен в приложении1.

4.3 Правила обеспечения косвенной конфиденциальности.

1. Классификационные признаки «Код основного вида экономической деятельности по ОКВЭД («хозяйственный ОКВЭД»)» «Код вида экономической деятельности по ОКВЭД («чистый ОКВЭД»)» относятся к сильно идентифицирующим переменным и подлежит агрегированию в соответствии с политикой в отношении распространения статистической информации (ОКВЭД1).

2. Классификационный признак «Код территории по ОКАТО» относится к предельно идентифицирующим переменным и подлежит агрегированию в соответствии с политикой в отношении распространения статистической информации:

Детализированный код ОКАТО, представленный в исходном массиве данных, в обязательном порядке агрегируется до первого уровня классификации, который включает объекты федерального значения (республики, края, области, города федерального значения, автономную область, автономный округ, входящий в состав Российской Федерации) и создается новый категориальный показатель ОКАТО1;

3. Классификационные признаки «Код формы собственности по ОКФС» и «Код организационно-правовой формы по ОКОПФ» относится к идентифицирующим переменным. Их рекомендуется удалить из массива микроданных для предотвращения излишней детализации данных, что позволит сохранить более глубокий уровень детализации данных по ОКВЭД, который является стратегически важным для научного анализа.

4. Показатели «Средняя численность работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации» и «Средняя численность работников территориально-обособленных подразделений, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера» относятся к сильно идентифицирующим переменным. Их значения необходимо агрегировать в несколько классов для того, чтобы снизить их идентифицирующие свойства, и создать новые переменные «Категория предприятия по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации» и «Категория предприятия по средней численности работников территориально-обособленных подразделений, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера». Границы классов приведены в приложении 2.

6. В отношении новых переменных «Категория предприятия по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации» и «Категория предприятия по средней численности работников территориально-обособленных подразделений, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера» следует применить процедуру оценки риска раскрытия в комбинации с кодами ОКВЭД1 (для статистических данных в разрезе по «хозяйственным» видам деятельности – строки 181, 200, 202 формы № 1-предприятие) или ОКВЭД2 (для статистических данных в разрезе по «чистым» видам деятельности – строки 182, 201, 203 формы № 1-предприятие) и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться методами защиты конфиденциальности снижения этого риска.

7. Показатель «Оборот организации (без НДС, акцизов и других аналогичных платежей) – всего по организации» относится к идентифицирующим переменным. В отношении данного показателя следует применить процедуру оценки риска раскрытия в комбинации с кодами ОКВЭД1 (для статистических данных в разрезе по «хозяйственным» видам деятельности – строки 181, 200, 202 формы № 1-предприятие) или ОКВЭД2

(для статистических данных в разрезе по «чистым» видам деятельности – строки 182, 201, 203 формы № 1-предприятие), «Категорией предприятия по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации», и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться методами защиты конфиденциальности для снижения этого риска.

8. Показатель «Отгружено товаров собственного производства, выполнено работ и услуг собственными силами» относится к идентифицирующим переменным. В отношении данного показателя следует применить процедуру оценки риска раскрытия в комбинации с кодом ОКВЭД1, «Категорией предприятия по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации» и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться неким методом защиты конфиденциальности для снижения этого риска.

9. Показатели «Оборот территориально-обособленного подразделения организации за отчетный год» и «Оборот территориально-обособленного подразделения организации за предыдущий год» относятся к идентифицирующим переменным. В отношении данных показателей следует применить процедуру оценки риска раскрытия в комбинации с кодами ОКВЭД1 (для статистических данных в разрезе по «хозяйственным» видам деятельности – строки 181, 200, 202 формы № 1-предприятие) или ОКВЭД2 (для статистических данных в разрезе по «чистым» видам деятельности – строки 182, 201, 203 формы № 1-предприятие), «Категорией предприятия по средней численности работников территориально-обособленного подразделения, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера», и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться методами защиты конфиденциальности для снижения этого риска.

10. Показатель «Фонд начисленной заработной платы работникам, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации» относится к идентифицирующим переменным. В отношении данного показателя следует применить процедуру оценки риска раскрытия в комбинации с кодами ОКВЭД1 (для статистических данных в разрезе по «хозяйственным» видам деятельности – строки 181, 200, 202 формы № 1-предприятие) или ОКВЭД2

(для статистических данных в разрезе по «чистым» видам деятельности – строки 182, 201, 203 формы № 1-предприятие), «Категорией предприятия по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера – всего по организации», и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться методами защиты конфиденциальности для снижения этого риска.

11. Показатель «Фонд начисленной заработной платы работникам территориально-обособленных подразделений, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера» относится к идентифицирующим переменным. В отношении данного показателя следует применить процедуру оценки риска раскрытия в комбинации с кодами ОКВЭД1 (для статистических данных в разрезе по «хозяйственным» видам деятельности – строки 181, 200, 202 формы № 1-предприятие) или ОКВЭД2 (для статистических данных в разрезе по «чистым» видам деятельности – строки 182, 201, 203 формы № 1-предприятие), «Категорией предприятия по средней численности работников территориально-обособленных подразделений, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера», и кодом ОКАТО1, если он сохранен в файле микроданных. После выявления объектов, подверженных риску ре-идентификации, если риск признается чрезмерным, следует воспользоваться методами защиты конфиденциальности для снижения этого риска.

12. По мере возможности сохраняются итоговые значения по показателям, соответствующим опубликованным итоговым данным, для обеспечения прозрачности.

5. Методы контроля обеспечения конфиденциальности данных после проведения деперсонификации

При выборе методов защиты конфиденциальных данных используются следующие критерии оценки их эффективности.

1. Безопасность. Метод должен обеспечивать недопустимость точной или очень близкой оценки значений переменных, что может привести к идентификации объекта. Уровень безопасности должен быть достаточно высоким.

2. Устойчивость. Метод считается устойчивым, если обеспечивается невозможность идентификации объекта даже в условиях наличия дополнительной информации об объектах из других источников.

3. Гибкость. Метод считается достаточно гибким, если он может быть использован одновременно для нескольких переменных, которые в свою очередь могут быть и численными и категориальными.

4. Полнота информации. Обработанные данные должны иметь высокую степень достоверности, т.е. незначительные, в пределах допустимых, отклонения в оценке статистических характеристик по сравнению с исходной совокупностью данных. Метод должен предоставлять возможность дальнейшей обработки данных: создание новых таблиц, проведение многовариантного анализа и т. п.

5. Стоимость. Учитывая возможные большие объемы обрабатываемой информации, стоимость применения метода должна быть разумной.

6. Простота. Процедура обработки данных должна быть понятной всем пользователям, не только профессиональным статистикам.

7. Инвариантность. Применяется по отношению к добавлению или удалению переменных или объектов из исходной совокупности, а также по отношению к линейным преобразованиям данных.

Приведенные выше критерии применяются для того, чтобы оценить и сравнить достоинства и недостатки различных методов защиты конфиденциальной статистической информации от полной или частичной идентификации объекта наблюдения.

Оценка эффективности применения методов защиты конфиденциальных статистических данных приведена в приложении 3.

6. Полезность и конфиденциальность. Показатели потери информации

Все методы защиты информации неизбежно ведут к снижению информационного содержания файла микроданных. При выборе оптимальных методов и решений для деперсонификации микроданных необходимо стремиться к минимизации риска раскрытия, обеспечивая при этом максимальную полезность статистических данных с позиции их последующего анализа.

Общий показатель потери информации оценивает объем потерянной информации применительно к разумному спектру способов использования данных. Потеря информации была небольшой, если массив защищенных данных является аналитически адекватным и интересным, в соответствии со следующими определениями:

- массив защищенных микроданных является аналитически адекватным, если следующие его характеристики примерно соответствуют оригинальным данным: средние значения и ковариации для отдельных подмножеств записей и/или переменных; характеристики распределения

- массив микроданных является аналитически интересным, если в нем представлены как минимум шесть переменных, пригодных для корректного анализа.

Оценка потери информации для численных переменных выполняется с использованием следующих описательных статистик, рассчитанных для оригинального и модифицированного (защищенного) массива данных:

- форма распределения переменных;
- средние значения;
- дисперсии;
- корреляции;
- ковариации;
- квантили переменных и соотношений переменных;

Сравнение должно быть выполнено для каждой комбинации ключевых категориальных переменных.

Пока нет единого количественного показателя, который полностью отражал бы структурные различия между оригинальными и защищенными данными, поэтому измерять потери информации и, соответственно, безопасность данных, предлагается также через различия между матрицей X для исходных данных и соответствующей матрицей X' для защищенного массива данных.

Величину расхождения (ошибки) между матрицами (X - X') можно измерять по крайней мере тремя способами.

- среднеквадратическая ошибка: сумма квадратов различий между соответствующими компонентами матриц, деленная на количество ячеек в каждой матрице:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-x_{ij}^{'}\right)^{2}}{np}$$

- средняя абсолютная ошибка: сумма абсолютных различий между соответствующими компонентами соответствующих матриц, деленная на количество ячеек в каждой матрице:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left|x_{ij}-x_{ij}^{'}\right|}{np}$$

- среднее отклонение: сумма абсолютных процентных отклонений компонент матрицы, рассчитанных для защищенных данных, от компонент матицы, рассчитанных для исходных данных, деленная на количество ячеек в каждой матрице. Преимущество данного подхода в том, что масштаб изменений переменных не имеет значения:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left|x_{ij}-x_{ij}^{'}\right|}{\left|x_{ij}\right|}}{np}$$

В вышеприведенных формулах p – количество переменных, n – количество записей, а компоненты матриц представлены соответствующими буквами в нижнем регистре (например, x_{ij} – компонента матрицы X). Деление на x_{ij} существенно увеличивает среднее отклонение X - X', если исходное значение x_{ij} близко к 0.

Поскольку такая зависимость от конкретного исходного значения нежелательна для показателя потери информации, предлагается заменить среднее отклонение X - X' на более стабильный показатель:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left|x_{ij}-x_{ij}^{'}\right|}{\sqrt{2}S_j}}{np},$$

где S_j - стандартное отклонение j-й переменной исходного массива данных.

Поскольку необходимо найти приемлемый баланс между потерей информации и риском раскрытия, а последний ограничен – не может быть риска выше 100%, - следует ввести верхнюю границу для показателя потери информации. На практике предлагается ограничить представленные выше показатели, основанные на среднем отклонении, неким заранее выбранным максимальным значением.

7. Описание процедуры деперсонификации данных годового структурного обследования предприятий

Формирование файлов общего пользования осуществляет Росстат.

Формализованное описание решения задачи деперсонификации данных годового структурного обследования предприятий может быть представлено в виде следующей последовательности действий:

анализ исходного файла данных;

предварительная обработка переменных;

оценка риска раскрытия;

выбор и применение методов деперсонификации;

оценка качества результата;

составление краткого описания изменений файла данных.

1) АНАЛИЗ ИСХОДНОГО ФАЙЛА ДАННЫХ
a) Формирование исходного массива микроданных годового структурного обследования по показателям в соответствии с утвержденным Росстатом перечнем.
b) Определение набора идентифицирующих переменных (прямых, косвенных).
c) Исключение из массива данных отдельных единиц наблюдения, доступ к данным которых не может быть предоставлен в соответствии с требованиями по защите государственной тайны.
2) ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ПЕРЕМЕННЫХ
a) Удаление прямых идентификаторов объектов, а также переменных, которые могут привести к спонтанной идентификации, либо к идентификации на основе сведений из внешних источников.
b) Перекодирование основных классификационных признаков по ОКВЭД, ОКАТО, ОКФС, ОКОПФ в соответствии с политикой в отношении публикации данных в этих разрезах.
c) Сокращение детализации отдельных сильно идентифицирующих численных переменных создание соответствующих новых категориальных переменных.
d) Анализ основных статистических характеристик сформированного массива исходных массива данных, изучение пользовательских предпочтений для последующего выбора методов деперсонификации и их параметров.
3) ОЦЕНКА РИСКА РАСКРЫТИЯ
a) Выявление небезопасных комбинаций ключевых категориальных переменных с использованием метода оценки риска на основе ключей (комбинации ключевых переменных, риск раскрытия которых необходимо оценить. Пороговое значение задается экспертами.
b) Выявление небезопасных комбинаций ключевых переменных, в составе которых есть численные переменные, с использованием методов оценки риска на основе алгоритмов кластеризации (комбинации ключевых переменных, риск раскрытия которых необходимо оценить. Значения параметров кластерного анализа задаются экспертами (приложение 3).
4) ВЫБОР И ПРИМЕНЕНИЕ МЕТОДОВ ДЕПЕРСОНИФИКАЦИИ

a) Если риск ре-идентификации на основе ключевых категориальных переменных признается чрезмерным – применение метода глобального перекодирования для соответствующих ключевых категориальных переменных. Далее – возврат на предыдущий шаг (процедура оценки риска на основе ключей).
b) Если риск ре-идентификации на основе ключевых категориальных переменных признается допустимым за исключением небольшого числа записей – принятие решения относительно ставшихся рискованных комбинаций ключевых переменных (варианты: сохранить в массиве данных «как есть»; применить перекодирование отдельных значений категориальных переменных, представленных в рискованных комбинациях; применить метод кодирования сверху и снизу).
c) Если риск ре-идентификации на основе ключевых численных переменных признается чрезмерным – применение методов модификации данных (варианты: обмен данными, микроагрегирование, добавление шума) к соответствующим численным ключевым переменным. Значения параметров методов модификации, определяющих уровень защиты данных и степень их искажения, устанавливаются экспертами. Далее – возврат на предыдущий шаг (процедура оценки риска на основе алгоритмов кластеризации).
d) Если риск ре-идентификации на основе ключевых численных переменных признается допустимым за исключением небольшого числа записей – принятие решения относительно значений численных ключевых переменных в оставшихся рискованных комбинациях (варианты: вменение значений ближайшей кластерной единицы, микроагрегирование только в хвостах; локальное подавление отдельных значений переменных, либо записи целиком).
e) При необходимости обеспечения соответствия ранее опубликованным данным – корректировка для сохранения суммарных значений отдельных показателей для каждой комбинации категориальных переменных, которые предполагается опубликовать.
f) При необходимости дополнительной защиты (в зависимости от формы распределения и статистических характеристик переменных) – применение метода округления для численных переменных, включенных в массив, но не признанных идентифицирующими.
5) ОЦЕНКА КАЧЕСТВА РЕЗУЛЬТАТА
a) Контроль обеспечения конфиденциальности данных после проведения деперсонификации с применением алгоритмов оценки риска.
b) Проверка предполагаемых к публикации переменных, которые могут привести к спонтанной идентификации, экспертами – специалистами по обследованию. Если таковые будут выявлены – использование индивидуальных методов защиты.
c) Оценка потери информации для численных переменных с использованием описательных статистик, рассчитанных для оригинального и модифицированного (защищенного) массива данных, в том числе: формы распределения переменных; средних значений; дисперсии; корреляции; ковариации; квантилей переменных и соотношений переменных и др.
6) СОСТАВЛЕНИЕ КРАТКОГО ОПИСАНИЯ ИЗМЕНЕНИЙ ФАЙЛА ДАННЫХ
a) Указать какие именно методы, к каким переменным были применены, но при этом без технических подробностей, которые позволили бы пользователям

восстановить идентифицирующие переменные.
b) Сообщить о степени модификации данных в результате использования методов деперсонификации (предоставить результаты оценки потери информации).

Схема процедуры формирования защищенного файла микроданных представлена на рис. 2.

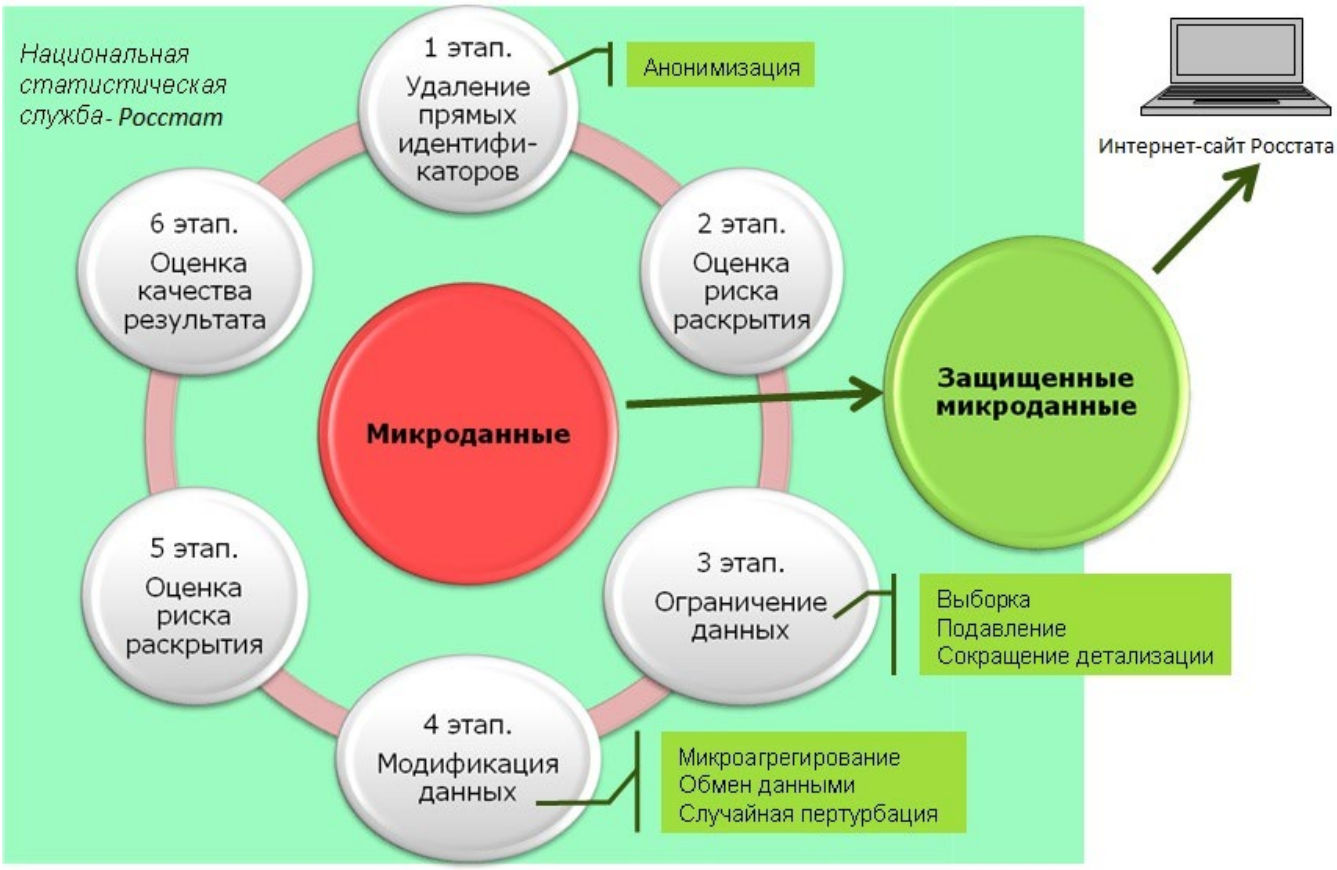

Рис. 2 – Формирование защищенного файла микроданных

На каждом этапе работа осуществляется экспертом с использованием специальных программных средств.

На 3 и 4 этапах осуществляется деперсонификация микроданных.

На 3 этапе используются методы, основанные на сокращении данных (непертурбативные методы); на 4 – методы, основанные на модификации данных (пертурбативные методы).

Непертурбативные методы применяются для непрерывных и категориальных данных:

Непертурбативные методы деперсонификации микроданных

Метод	*Непрерывные данные*	*Категориальные данные*
Формальная анонимизация (обезличивание)		X
Выборка		X
Сокращение детализации	X	X
Локальное подавление	X	X

В качестве методов модификации микроданных с целью их маскировки как наиболее простые и обеспечивающие необходимый уровень качества модифицированного массива данных определены следующие:

- метод обмена данными
- метод микроагрегирования
- случайная пертурбация.

Пертурбативные методы деперсонификации микроданных

Метод	*Непрерывные данные*	*Категориальные данные*
Обмен данными	*X*	*X*
Микроагрегирование	*X*	
Случайная пертурбация (добавление шума)	*X*	

Для достижения оптимального соотношения между риском и полезностью данных процедуру модификации данных рекомендуется выполнять в несколько итераций, постепенно усиливая степень модификации данных и выполняя оценку качества результата на каждом шаге, с целью достижения максимально допустимого уровня модификации данных, на котором следует остановиться.

Заключение

Обеспечение потребности широкого круга исследователей в статистических данных в максимально детализированной форме – микроданных для научных, аналитических и других целей ставит перед органами статистики задачу создания массивов деперсонифицированных микроданных. С целью реализации возможности работы пользователей с микроданными можно предложить два основных пути:

– формирование файлов микроданных общего пользования для размещения их в открытом доступе;

– формирование файлов научного пользования по персональным запросам.

Файлы общего пользования предназначены для использования широким кругом пользователей и размещаются в открытом доступе (на сайте Росстата). Исходя из этого они содержат абсолютно анонимные данные, риск раскрытия конфиденциальной информации минимизируется путем соответствующей структуризации этих файлов с помощью методов контроля раскрытия статистической информации, отвечающих самым строгим требованиям надежности. При создании файлов общего пользования требуется проведение экспертных процедур, которые носят итерационный характер.

ПРИЛОЖЕНИЕ 1

Перечень переменных формы № 1-предприятие «Основные сведения о деятельности организации» не подлежащих включению в файл микроданных

■Наименование отчитывающейся организации,

Почтовый адрес

Код отчитывающейся организации (по ОКПО) и Код вышестоящей организации (по ОКПО)

Код формы собственности по ОКФС

Код организационно-правовой формы по ОКОПФ

Показатели, характеризующие демографию предприятия (Раздел I. Общие сведения о юридическом лице);

■ Показатели, характеризующие организационную структуру юридического лица (Раздел II. Сведения об изменениях юридического лица в отчетном году, Раздел V. Организационная структура юридического лица в отчетном году);

■ Показатели, характеризующие уставный капитал организации (Раздел III. Распределение уставного капитала (фонда) между акционерами (учредителями), Раздел IV. Взносы иностранных юридических и физических лиц в уставный капитал (фонд) по странам – партнерам).

Приложение 2

Границы классов
для определения категории предприятия территориально-обособленных подразделений по средней численности работников, включая внешних совместителей и работников, выполнявших работы по договорам гражданско-правового характера

1) до 50 чел.
2) 50–99 чел.
3) 100–199 чел.
4) 200–249 чел.
5) 250–499 чел.
6) 500–999 чел.
7) 1000–4999 чел.
8) 5000–9999 чел.
9) 10 000 и более чел.

Приложение 3

**Первоначальные значения параметров,
используемых в алгоритмах деперсонификации
на примере годового структурного обследования**

Наименование параметра	Значение
Минимальное количество предприятий в группе	3
Шаг кластеризации для показателя "оборот"	500000
Шаг кластеризации для показателя "отгружено"	10000
Шаг кластеризации для показателя "фонд ЗП"	10000
Параметр для метода "Обмен данными" (%)	5
Шаг кластеризации для показателя "оборот ТОП"	500000
Шаг кластеризации для показателя "фонд ЗП ТОП"	10000
Количество знаков ОКВЭД для укрупнения	3
Количество знаков ОКВЭД для укрупнения кодов 50,52	5

Приложение 4

Оценка эффективности применения отдельных методов защиты конфиденциальных статистических данных

Критерий		Обмен данными	Микроагрегирование					Добавление шума
			Ранжирование по одной переменной	Ранжирование по первой главной компоненте	Ранжирование по сумме нормированных величин	Индивидуальное ранжирование	Индивидуальное ранжирование со взвешенным скользящим средним	
1. Безопасность		высокая	средняя	высокая	высокая	высокая	высокая	высокая
	а) точное раскрытие	нет	нет	нет	нет	нет	нет	нет
	б) частичное раскрытие	затруднено	да, возможно	затруднено	затруднено	затруднено	затруднено	да, зависит от «шума»
2. Устойчивость		высокая	низкая	средняя	средняя	средняя	средняя	средняя
3. Гибкость		средняя	средняя	средняя	средняя	высокая	высокая	низкая
4. Полнота информации		от средней к низкой	низкая	средняя	средняя	высокая	высокая	средняя
	а) потери информации	от низких до средних	высокие	средние	средние	низкие	от низких до средних	средние
	б) полезность для анализа	зависит от параметров	низкая	средняя	средняя	высокая	от средней до высокой	от низкой до средней
5. Стоимость		высокая	низкая	низкая	низкая	низкая	низкая	низкая
6. Простота	от 1- очень просто до 5 - очень сложно	5	1	3	2	1	2	2
7. Инвариантность		да	да	нет	нет	да	да	да