ТИПОВАЯ МОДЕЛЬ СТАТИСТИЧЕСКОГО РЕДАКТИРОВАНИЯ ДАННЫХ

2021

Содержание

От автора

Я проработал в органах статистики разных стран мира в общей сложности 35 лет. Над вопросами оптимизации и эффективности процесса редактирования статистических данных я размышлял, в том числе совместно со своими коллегами, на протяжении нескольких десятков лет. Эта тематика актуальна и для других статистических служб мира.

В 2019 году коллективом специалистов была разработана Типовая модель редактирования статистических данных. Представленный далее материал является адаптированной и переработанной версией данной модели, которая может быть использована как статистическими службами, так и специалистами, деятельность которых связана с редактированием статистических данных.

Типовая модель статистического редактирования представляет собой изложение стандартизированного подхода к процессу редактирования и импутации статистических данных. Она представлена в виде справочного руководства по редактированию и импутации статистических данных и направлена на улучшение понимания, коммуникации, практического применения, развития в области редактирования статистических данных.

Данная модель с одной стороны, как формализованное и достаточно обобщенное описание существующей практики, а с другой как некий эталон или стандарт, позволяет эксперту оценить используемые на практике техники и методики с точки зрения соответствия их международным стандартам в области редактирования, и на основе проведенного анализа выработать план действий, направленный на повышение качества получаемых статистических данных.

Алексей Хохловский, Санкт-Петербург

1.Аннотация

Тема редактирования данных привлекает значительный интерес в контексте модернизации официальной статистики, поскольку оно традиционно является одним из самых дорогостоящих и трудоемких частей процесса статистического производства и подвержено влиянию таких инновационных процедур, как машинное обучение. Поэтому обмен идеями, опытом и передовой практикой для повышения эффективности редактирования данных наряду с четко определенными и хорошо описанными процедурами редактирования является приоритетной задачей для международного статистического сообщества.

Типовая модель редактирования статистических данных (TMPC) была разработана под руководством Группы высокого уровня по модернизации официальной статистики (HLG-MOS). Она предназначена для использования в качестве справочного материала всеми статистиками, деятельность которых предусматривает редактирование данных.

Ручное или интерактивное редактирование, выполняемое профильными специалистами, требовало значительных временных и материальных затрат и отрицательно влияло на своевременность публикаций. Повторное обращение к респондентам в процессе интерактивного редактирования приводило к увеличению нагрузки на респондентов.

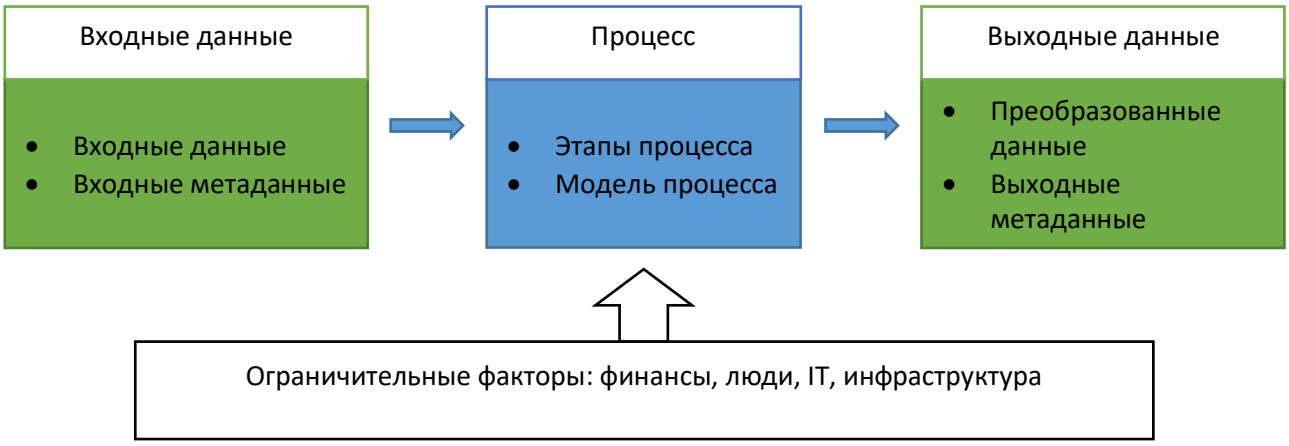
Многочисленные исследования в этой области показали, что количество записей, редактируемых вручную, можно значительно сократить, если усилия по редактированию будут сосредоточены на ошибках, оказывающих наибольшее влияние на оценки основных представляющих интерес параметров.

Модель может использоваться в качестве стандартного справочного материала для редактирования статистических данных, точно так же, с методологической точки зрения, как и набор стандартных моделей и методов для оценки результатов обследований.

2.Процесс редактирования статистических данных

Процесс редактирования статистических данных представлен на рисунке 1. Данные и метаданные рассматриваются как входные данные, ряд действий выполняется для оценки достоверности данных, выявления потенциальных проблем и их устранения; преобразованные данные – выходные данные. Процесс находится в окружении ограничивающих факторов, как показано на рисунке 1.

Рисунок 1 Процесс статистического редактирования



Редактирование статистических данных часто также называют редактированием и импутацией. На протяжении всего документа будет в основном использоваться первый термин (а также "редактирование данных" или просто "редактирование"). При этом в случаях, когда речь заходит о широко используемых выражениях, таких как, например, "первоначальное редактирование и импутация " будет использоваться последний.

Процесс редактирования данных в основном состоит из бизнес-функций, в рамках которых выполняются определенные задачи с определенными целями. В контексте редактирования статистических данных мы называем эти функции "функциями редактирования данных" или "функциями "для краткости. С точки зрения их назначения эти функции можно разделить на три типа функций: "проверка ", "отбор" и "обработка ", основываясь на терминах, предложенных Паннекоеком и Чжаном (2012). Эти типы функций также можно рассматривать как высокоуровневые функции, тем самым характеризуя процесс редактирования данных с точки зрения выполнение трех задач: проверка, отбор и/или обработка.

Редактирование статистических данных включает или оказывает влияние на все восемь этапов Типовой модели статистического производства (GSBPM). Согласно GSBPM статистическое производство можно рассматривать как процесс преобразования исходных входных данных в статистическую информацию. Редактирование данных является частью этого производственного процесса. Иногда все операции редактирования данных могут быть сгруппированы таким образом, чтобы сформировать некий "фиксированный сегмент" в цепочке с одной точкой входа и одной точкой выхода (например, в рамках этапов GSBPM 5.3 "Проверка и валидация " и 5.4 "Редактирование и импутация ").

Однако в целом редактирование данных может применяться в различных областях в течение жизненного цикла данных, в том числе в случае, когда ранее обработанные данные повторно используются или/и объединяются с другими данными для получения новых статистических результатов, таких как редактирование национальных счетов или других макрорасчетов.

Например, взвешивание выборочных единиц - это процесс, реализуемый в рамках этапа 5 модели GSBPM. По общему правилу взвешивание не считается редактированием данных, хотя оно, по сути, является статистической функцией имеющей отношение к редактированию как входных, так и выходных данных.

С этой точки зрения редактирование во время сбора данных (этап 4 GSBPM), в том числе в при применении инструментов сбора, представляет собой либо процесс редактирования данных, либо подэтап, в зависимости от объема процесса и интерпретации входных и выходных данных. Традиционно существуют также споры по поводу импутации с точки зрения редактирования и импутации с точки зрения оценки. Кроме того, уточняя цели и использование функций обработки в процессе редактирования данных, мы ,будем или, наоборот, не будем уделять особое внимание конкретному процессу импутации как одной из частей процесса редактирования данных и достигать соглашения по этим вопросам.

В этом документе особое внимание уделяется главным образом внедрению процессов редактирования данных. Планирование, разработка и оценка стратегий редактирования, как правило, рассматриваются вне сферы охвата (этого документа), хотя при этом приводятся ссылки на показатели процесса (параданные), поскольку эта информация может использоваться непосредственно в процессе редактирования.

Описание функций редактирования данных предоставляет средства для концентрации на сути этой структуры, в то время как ориентация на обслуживание помогает обеспечить ее охват для того, чтобы сделать структуру достаточно универсальной.

Наконец, настоящий документ в первую очередь ориентирован на обработку данных для достижения пригодности к использованию. Другие важные цели редактирования статистических данных, такие как оценка качества и предотвращение будущих ошибок, могут основываться на результатах различных функций редактирования данных, таких как проверка и отбор. В данном документе они не детализированы и не проработаны.

3.Используемые термины

Общий процесс редактирования статистических данных, описанный в настоящем документе, предусматривает проведение ряда мероприятий или выполнение задач, направленных на оценку достоверности данных, выявление потенциальных проблем и выполнение отдельных действий для устранения выявленных проблем.

Ниже приводятся предложения по определению и описанию некоторых основных элементов, которые можно выделить в процессе редактирования данных, а также входных и выходных данных этого процесса. Эти предложения основаны на более общих определениях Типовой модели статистической информации (GSIM), которые применяются к конкретному контексту редактирования данных. Используются следующие определения:

- **Функции и методы. В этом разделе описываются элементы, связанные с выполнением различных задач редактирования данных.**
- **Типы метаданных. В этом разделе описываются метаданные, необходимые для определения и описания процесса редактирования данных, а также формирование выходных данных в рамках этого процесса.**
- **Технологический поток, этап процесса и контроль. В этом разделе описываются элементы, связанные с организацией процесса редактирования данных.**

4.Функции и методы

GSIM рассматривает бизнес-функцию как "что-то, что предприятие делает или должно делать для достижения своих целей". Функция редактирования статистических данных - это бизнес-функция, при реализации которой преследуют определенную цель в цепочке действий, определяющих процесс редактирования данных.

Они могут подразделяться на три широких категории:

- **Проверка. Функции, в ходе реализации которых исследуют данные для выявления потенциальных проблем.**
- **Отбор. Функции, в ходе реализации которых выбирают единицы или их части для дальнейшей обработки.**
- **Обработка. Функции, в ходе реализации которых данные преобразуют таким образом, чтобы улучшить их качество. Преобразование единиц (т. е. заполнение пропущенных значений или изменение ошибочных) называется импутацией.**

Наиболее распространенными примерами функций, относящихся к одной из трех категорий, являются:

- **Проверка: оценка правдоподобности значений или их комбинаций; оценка данных на предмет логической согласованности; измерение правдоподобности оценок на макроуровне.**
- **Отбор: выбор единиц для интерактивной обработки; отбор единиц – выбросов для отдельной обработки; отбор влиятельных (значимых) единиц – выбросов для отдельной обработки; отбор переменных для обработки с помощью конкретных методов импутации; определение ошибочных значений среди тех, которые признаются противоречивыми.**
- **Обработка: импутация отсутствующих или отброшенных (ошибочных) значений; исправление систематических ошибок; корректировка несоответствий.**

Различные типы функций часто взаимосвязаны и могут быть упорядочены следующим образом: функции проверки с использованием показателей качества или его измерениям, которые указывают на конкретные проблемы в данных; функции отбора с использованием показателей качества и/или критериев отбора (пороговых значений), а также входных данных, и позволяют сформировать индикаторы, идентифицирующие записи или их части для дальнейшей обработки; наконец, функции обработки предполагает изменение или импутацию выбранных значений данных для решения

обнаруженных ранее проблем. Затем результаты могут быть подвергнуты другой (или следующей) проверке.

К каждому типу функции относятся основные типы входных и выходных данных и метаданных. Входные данные используются во всех функциях, в то время, как только при реализации функции обработки производятся новые данные в качестве выходных. Входные и выходные метаданные для каждого типа функций рассматриваются в главе 4.

5.Методы

Функции редактирования данных определяют, какое действие должно быть выполнено, но не то, каким образом оно выполняется. Последнее определяется используемым методом. Примерами методов для различных типов функций являются:

- Проверка: оценка в соответствии с функцией оценки или счетчика, применение набора правил редактирования; проведение расчетов для обнаружения выбросов.
- Отбор: использование заданного критерия для определения выбросов; использование заданного порога для конкретной функции оценки (счетчика) для селективного редактирования; отбор единиц (определенного их процента) с наибольшими значениями баллов; применение парадигмы Филледжи - Хольта для определения ошибок с определенными весами.
- Обработка: конкретные методы и модели импутации для заданных переменных; обеспечение согласованности конкретных переменных согласно конкретным алгоритмам; корректировка значений предметными специалистами.

6.Типы метаданных

В процессе редактирования в качестве вводных данных используются входные данные и входные метаданные. Входные данные-это данные, которые являются объектом редактирования. Входные метаданные представляют собой информацию, необходимую для запуска процесса. На стороне выхода находятся преобразованные данные, которые соответствуют входным данным (с модификациями) и выходные метаданные, содержащим дополнительную информацию, полученную в результате процесса редактирования.

Согласно рисунку 1, метаданные, необходимые для выполнения редактирования данных, можно в широком смысле классифицировать следующим образом:

1)Метаданные на этапе входа. Информационные объекты, описывающие процесс редактирования статистических данных на входе. Входные метаданные включают концептуальные и структурные элементы метаданных, полезные с точки зрения описания входных данных (единицы, переменные, разрезность, набор данных, записи...) и дополнительную информацию,

необходимую для реализации функций, например вспомогательные данные и параметры.

2)Метаданные этапов процесса и потока (модели). Информационные объекты, необходимые для описания самого процесса редактирования статистических данных. Каждый этап процесса детализирован с точки зрения функций и методов, в то время как последовательность между этапами процесса регулируется с помощью инструментов контроля процесса.

3)Выходные метаданные. Информационные объекты, описывающие выходные данные в процессе редактирования статистических данных. Выходные метаданные процесса включают концептуальные и структурные элементы метаданных, полезные с точки зрения описания выходных данных. Другие метаданные, полученные в процессе редактирования, представляют собой информацию о качестве, используемую как для входных, так и для выходных данных. Кроме того, может быть собрана информация о ходе выполнения процесса, которая не имеет прямого отношения к качеству данных (параданные).

7.Этап процесса, процесс и контроль

Элементы, необходимые для разработки и описания конкретного процесса редактирования статистических данных, включают следующие элементы.

Этап процесса

Рабочий процесс редактирования данных обычно содержит значительное количество функций с заданными методами, которые выполняются определенным образом. Чтобы описать организацию всего процесса в понятной форме, полезно разделить процесс на ограниченное число этапов процесса и описать всю организацию в терминах этапов процесса.

Процесс и контроль

Описание процесса в терминах этапов процесса должно также включать описание последовательности перехода между ними. Поток процесса включает реализуемые этапы процесса и последовательность их выполнения.

Обычная последовательность - это когда за одним этапом следует один и тот же этап при любых обстоятельствах. Когда за этапом может следовать несколько альтернативных шагов, в зависимости от набора условий, то эта развилка управляется специальным элементом потока, который называется «контролем», описывающим разветвление в последовательности процесса.

Примеры этапов процесса высокого уровня:

- начальное редактирование и импутация (или редактирование отдельной области и редактирования систематических ошибок);
- автоматическое редактирование и импутация;
- интерактивное редактирование и импутация;
- макро редактирование и импутация;
- объединение и согласование.

Примеры элементов контроля:

-отбор единиц со значимыми подозрительными значениями для интерактивной обработки;
- отбор переменных для специальной обработки (например, импутация каким-либо подходящим методом, методы редактирования категориальных/непрерывных переменных);
- поиск глубинных причин возникновения подозрительных агрегатов.

8.Функции и методы

Введение

Функции и методы являются существенной частью описания более низких уровней иерархии процесса редактирования данных. В этой главе представлена более точная классификация и даны определения функций и методов, используемых в процессах редактирования, а также приводятся примеры и комментарии. В документе использованы несколько модифицированные понятия и структуры, которые ранее были описаны у Camstra и Renssen (2011), Pannekoek and Zhang (2012) и Pannekoek и др. (2013).

Различие между функциями и методами можно понять следующим образом. Функции представляют собой набор действий по редактированию данных, которые должны выполняться, методы – каким образом эти действия должны выполняться. Функция может быть реализована несколькими методами и один метод может выполняться с использованием различных функций.

9.Функции

Функция редактирования статистических данных-это бизнес-функция, которая реализуется с определенной целью в установленной последовательности, определяющей процесс редактирования данных. Функции могут быть подразделены на три широких категории:

- Проверка. Функции, которые предусматривают исследование данных для выявления потенциальных проблем.

- Отбор. Функции, которые предусматривают отбор единиц или значений показателей, относящихся к данным единицам, для дальнейшей обработки.

- Обработка. Функции, которые предусматривают изменение данных таким образом, чтобы улучшить их качество. Изменение значений показателей, относящихся к единицам (т. е. восполнение пропущенных или коррекция ошибочных значений) называется импутацией.

Функции могут быть дополнительно классифицированы в зависимости от задач, для реализации которых они выполняются, вида получаемого результата и от того по отношению к чему они применяются - к единицам или переменным. Описание категорий функций выглядит следующим образом. В таблице 1 приведены примеры этих функций.

- проверка достоверности данных (путем проверки комбинаций значений). Функции, которые предусматривают проверку валидности комбинации значений данных по сравнению с заданным диапазоном или набором значений, а также валидности заданных комбинаций значений. Каждая проверка приводит к бинарному результату (правда - ложь).

- проверка достоверности данных (путем анализа). Функции, которые предусматривают вычисление показателей правдоподобия значений данных в наборе данных (комбинации единиц). Это приводит к вычислению количественных показателей, которые могут быть использованы для оценки достоверности значений данных, в том числе агрегатов. Это также включает в себя менее формально определенные "функции", такие как анализ путем проверки графических изображений.

- проверка единиц. Функции, которые предусматривают вычисление баллов/очков, являющимися показателями качества для последующего отбора единиц. Функция оценки баллов/очков может быть какой угодно, описывающей единицу. Результат функции оценки баллов/очков часто необходим для дальнейшего использования на следующем этапе процесса, в котором выходной результат рассматривается как входные данные.

- отбор единиц. Функции, которые предусматривают отбор единиц из набора данных для отдельной обработки. Автоматический отбор используется, например, при сравнении значений функций оценки баллов/очков с заранее установленным пороговым значением. Соответственно, ручной отбор обычно основывается на макроредактировании, например, с использованием агрегатов и графики.

- отбор переменных. Функции, которые позволяют определить переменные, относящиеся к единицам измерения, для иной обработки, чем остальные переменные, с учетом наблюдаемых (предполагаемых) ошибок. Что касается единиц, то эта операция может быть выполнена либо вручную (проверка сотрудником), либо автоматически (обнаружение ошибок измерения, метод Филледжи-Хольта для определения ошибок)

- импутация переменных. Функции, которые предусматривают изменение наблюдаемых значений или заполнение пропущенных значений для улучшения качества данных. Обычно функция импутации предназначены для исправления различных типов ошибок (например, систематических ошибок, ошибок в характеристиках единиц измерения). Эти функции могут выполняться как автоматически (множество различных методов) так и

вручную (например, интерактивные операции). Под импутацией понимается как вменение отсутствующих значений полей, так и изменение ошибочных.

- обработка единиц. Функции, которые предусматривают изменение структуры единицы путем объединения (т. е. слияния) и согласования (обеспечения соответствия) различных единиц, относящихся к разным источникам. Цель состоит в том, чтобы получить и отредактировать целевые статистические единицы, которые не были сформированы заранее.

Таблица 1

Тип функции	Категории	Примеры
Проверка	Проверка достоверности данных (путем проверки комбинации значений)	1)Проверка на наличие очевидных ошибок. 2)Оценка логической согласованности комбинации значений. 3)Проверка свойств данных.
	Проверка правдоподобия данных (путем анализа)	1) Оценка правдоподобия значений или их комбинаций. 2) Оценка правдоподобия на макро-уровне. 3)Проверка и идентификация систематических ошибок. 4) Проверка объединения единиц на макро-уровне.
	Проверка единиц	1) Проверка единиц, соответствующих критериям. 2) Проверка единиц, не соответствующим критериям. 3) Проверка путем подсчета очков/баллов для значимых единиц и выбросов. 4) Проверка согласованности единиц на микро-уровне.
Отбор	Отбор единиц	1) Отбор единиц, соответствующих критериям. 2) Отбор единиц для интерактивной обработки, для неинтерактивной обработки, тех, что не подлежат обработке. 3) Отбор единиц, подвергающихся влиянию значимых ошибок. 4) Отбор выбросов для обработки с применением весовой корректировки. 5) Отбор по структуре единиц. 6) Отбор единиц после проверки на макро-уровне.

	Отбор показателей	1) Отбор переменных с очевидными ошибками. 2) Отбор переменных с ошибками в свойствах единиц. 3) Отбор переменных для обработки специфическими методами импутации. 4) Отбор влиятельных выбросов для ручной обработки. 5) Определение некорректных значений среди несогласованных/ противоречивых значений. 6) Определение показателей, находящиеся под влиянием ошибок для каждой единицы
Обработка	Импутация показателей	1) Коррекция очевидных ошибок. 2) Коррекция систематических ошибок. 3) Коррекция ошибок в свойствах единиц. 4) Импутация в случае ошибок. 5) Импутация пропущенных/отбракованных (ошибочных) значений. 6) Коррекция в случае несогласованности значений.
	Обработка единиц	1)Обработка единиц в критическом наборе. 2)Создание статистических единиц. 3)Работа с ошибками при объединении.

Помимо набора данных, который пересматривается и/или модифицируется, для реализации функций необходимы дополнительные метаданные для включения в поток процесса и этапы процесса.

Эти метаданные можно классифицировать следующим образом: входные метаданные процесса, метаданные для реализации функций, выходные метаданные процесса. Входные метаданные включают вспомогательные данные, параметры и неструктурированные метаданные.

В метаданных для реализации функций определены методы и правила. Выходные метаданные - это показатели качества и параданные.

10. Методы

Методы устанавливают, каким образом, должны выполняться функции редактирования данных в реальных жизненных ситуациях. В этом документе они называются методами для краткости.

Методы могут быть связаны с правилами.

Правила - это математические/логические функции в наборе данных и, возможно, также вспомогательных переменных. Иногда эти правила могут быть дополнительно детализированы с помощью параметров. Процесс определения корректных параметров в определенном контексте называется параметризацией. Мы различаем правила редактирования, функции счета (оценки) и правила редактирования.

Правила редактирования описывают допустимые (жесткие правила) или вероятные (мягкие правила) значения переменных или их комбинаций. Особенно в бизнес-статистике часто существуют большие наборы жестких и мягких правил редактирования, таких как линейные равенства (балансовые равенства), неравенства и контроль соотношений.

Правила редактирования используются в функциях проверки, которые оценивают нарушение жестких контролей или количество нарушений мягких контролей. Жесткие правила редактирования также используются для отбора значений, предположительно ошибочных, например с помощью метода Филледжи-Хольта. Методы импутации также могут предусматривать использование правила редактирования, в частности, корректировка для достижения согласованности импутированных значений предполагает использование жестких правил редактирования.

Функции счета (балльная оценка) позволяют оценить правдоподобие и влияние значений единицы в целом. Они обычно используются при реализации функции отбора, предусматривающей выбор единиц для интерактивного редактирования.

Правила редактирования сочетают в себе обнаружение, отбор и импутацию недостающих данных или ошибочных значений. В частности, в случае конкретных "очевидных" ошибок они используются для исправления систематических ошибок или, в более широком смысле, ошибок с обнаруживаемой причиной и известным механизмом появления ошибок. Они могут быть сформулированы как правила типа IF-THEN следующего вида: IF (условие) THEN OldValue (старое значение) = NewValue (новое значение). Этот тип правил обычно применяется во время микроредактирования. Правила типа IF-THEN также можно использовать для автоматического обнаружения ошибок. Они могут быть выражены в форме IF-THEN следующим образом: IF (условие) THEN Flag Value (значение флажка)= error Code (код ошибки).

Таблица 2, Таблица 3 и Таблица 4 имеют аналогичную структуру: каждая функция соотносится с одной или несколькими категориями методов с примерами. Некоторые из методов также присутствуют в общих этапах процесса, описанных в таблице 6 в главе 5. Классификация подкатегорий не включает все возможные альтернаты, хотя она демонстрирует много известных (широко используемых) методов для каждого из трех типов функций.

11.Проверка

Методы для реализации функции проверки варьируются от простых до сложных. Наиболее распространенными методами проверки являются различные правила редактирования. Методы, направленные на изучение

достоверности/правдоподобия данных, обычно требуют специальных аналитических инструментов для получения индикаторов отбора.

Оценка (балльная оценка) - это показатель качества единицы. Проверка с применением расчета баллов единицы включает две основные части: оценки (баллы) для селективного редактирования и другие типы расчетов для проведения проверки. Согласованность на микроуровне предусматривает выявление проблемных единичных ситуаций, связанных с объединением и обеспечением взаимосвязи/согласованности между несколькими входными источниками данных. В таблице 2 приведены примеры этих методов проверки.

Таблица 2

Функция	Категории методов	Примеры
Проверка достоверности данных	Правила редактирования	1) Редактирование посредством определения допустимых значений (набор допустимых значений, для переменной).
		2) Правила редактирования предусматривают ограничения (интервал допустимых значений для переменной)
		3)Правила редактирования предусматривают исторические сравнения (отношения значений переменных в разных точках времени)
		4) Правила редактирования предусматривают учет отношений переменных (построение отношений переменных на основе предварительных знаний)
		5)Комбинация типов правил редактирования (сочетание различных правил редактирования)
Проверка достоверности данных	Аналитические методы проверки	1)Критерии для обнаружения выбросов (например, вычисление значений на основе распределения переменной).
		2) Агрегаты для исследования на макро-уровне (например, вычисление итогов для сравнения с предыдущими итогами)
		3)Анализ охвата (например, содержит ли подсовокупность большую долю не совпадающих единиц?)
		4) Размер генеральной совокупности (например, число из регистра домашних хозяйств ≈ числу домашних хозяйств в переписи населения?)
		5) Кластерный анализ (определение ошибочных значений при смешанном моделировании).
Проверка единиц	Исследование состоятельности единицы	Проверка состоятельности значений показателей, относящихся к единице (проверка с точки зрения неответов и значений показателей, относящихся к единице)
	Согласованность на микро-уровне	Правила редактирования предусматривают проверку связей (например, проверка статуса соответствия, не совпадений, множества

		совпадений)
		Правила редактирования несоответствий (например, у человека несколько адресов?)
	Балльная оценка (счетчик) по вспомогательным переменным	Вспомогательная переменная как критерий значимости (например, использование оборота для оценки значимости предприятия)
	Расчет балльной оценки (счетчика) для селективного редактирования	Расчет балльной оценки (счетчика) для итоговых значений (количественная оценка влияния редактирования единицы на оценку итогов)
		Расчет балльной оценки (счетчика) для данных с ошибками(параметрическая модель, которая учитывает возможные)
		Расчет балльной оценки (счетчика) с учетом правил редактирования (вычисление счетчика (балльной оценки) с учетом правил редактирования и оценок)
		Расчет балльной оценки (счетчика) посредством латентного анализа классов (вычисление счетчика на основе ожидаемой ошибки на базе моделирования)
		Расчет балльной оценки посредством модели прогнозирования (прогнозирование вероятности ошибки на основе хорошо отредактированных данных предыдущего периода)
	Интерактивная проверка единиц	Проверка единицы и значений переменной в целом (оценка состояния единицы)

12.Отбор

Отбор приводит к простому результату - либо единица/переменная, относящаяся к единице, помечается как выбранная, либо нет (бинарность, 0/1). Используются пороговые значения, автоматическое тестирование структуры единицы либо ручной отбор, основанный на решениях лица, ответственного за редактирования. Соответственно, применяются различные вычислительные решения для ограничения набора переменных в наблюдениях для дальнейшей обработки. Опять же, ручной отбор -это один из способов решения проблемы. Правила для агрегатов могут напоминать принципы редактирования наблюдений. В таблице 3 представлены некоторые методы, реализующие теоретические типы отбора, а также практические решения.

Таблица 3

Функция	***Методы***	***Примеры***

Отбор единиц	Отбор по балльной оценке (значению счетчика)	Отбор по фиксированному порогу (используется порог, определяемый на основе опыта и доводах)
		Отбор по пороговому значению на основе расчета распределения баллов (точка из распределения используется в качестве порога)
		Отбор по пороговому значению на основе псевдо-смещенного исследования (процент уровня ручной обработки при псевдо-смещенном исследовании используется для определения порогового значения)
	Отбор по структуре	Сложные взаимосвязи (например, не состоящая в браке пара с ребенком проживает по одному адресу и жена мужа - по отдельному адресу)
		Подозрительная структура (например, адрес семейной ячейки - двоюродная бабушка и посторонний человек)
	Отбор на макро- уровне	Отбор по статистике группы (например, почтовые индексы с высокой вероятностью ошибки объединения)
	Интерактивной отбор единицы	Интерактивно отобранные единицы (технический отбор единиц)
Отбор переменных	Отбор переменных на микро-уровне	Отбор очевидных ошибок (направление очевидных ошибок на исправление с отбором)
		Определение случайных ошибок (определение ошибочного значения с помощью алгоритма)
		Принятие ситуации многовариантной ошибки, связанной с единицей (отбор всех переменных с индикатором, относящемуся к единице)
	Отбор переменных на макро уровня	Отбор, основанный на вычислении выбросов (правила отбора выбросов, основанные на специальном методе)
		Отбор, основанный на правилах для агрегатов (выявление подозрительного набора единиц на основе оценки)
	Отбор интерактивной переменной	Интерактивно отобранные переменные (отбор переменной специалистом для дальнейшей обработки)

13. Обработка

Функция обработки обычно располагает большим количеством соответствующих альтернативных методов, которые знакомы из литературы, а также практикой редактирования в процессах статистического редактирования. Некоторые методы могут подразделяться на более общие классы, например методы импутации переменных - на случайные и неслучайные. Обработка на уровне единиц обычно связана с различными операциями, выполняемыми при объединении и согласовании различных единиц, находящихся в нескольких входных источниках. В таблице ниже представлено описание несколько методов.

Таблица 4

Функция	Методы	Примеры
Импутация переменной	Интерактивная обработка ошибок	Повторно связаться (запрос реального значения стоимости от респондента или от поставщика данных).
		Проверка вопросников (проверка значения из вопросника, например, для обработки ошибок).
		Замена значения (замена или добавление значения из другого источника или от другой переменной).
		Создание значения (принятие решения в отношении значения, основанному на знании существа вопроса).
	Дедуктивная импутация	Импутация – функция (значение рассчитывается как функция других значений).
		Импутация на основе логических выводов (вычисление значения на основе логических выражений).
		Импутация с помощью исторических значений (значения, используемые из данных, относящихся к более ранним периодам времени).
		Прокси импутация (значение, полученное от соответствующей единице).
	Импутация на основе модели	Импутация среднего значения (использование среднего значения переменной). Импутация медианы (использование медианного значения переменной).
		Импутация показателя (использование значения вспомогательной переменной посредством расчета соотношений)
		Импутация на основе регрессии (прогнозирование значения с помощью регрессивной модели).
	Импутация донора	Импутация случайного донора (отбор донора случайным образом в пределах области). Последовательная импутация донора (последовательный отбор доноров).
		Импутация ближайшего соседа (отбор доноров, основанный на функции расстояния). Импутация ближайшего соседа случайным образом (отбор доноров случайным образом по соседству).
	Обеспечение согласованности	Решение по балансу (решение как результат, полученный исходя из условий согласованности) Пропорциональное распределение (корректировка существующих значений для обеспечения согласованности)
		Импутация донора с поправочным коэффициентом (импутация доноров с поправочным коэффициентом в целях согласованности)
		Частичная корректировка переменных (коррекция значений переменных на основе предварительных

		знаний)
Правка единицы	Отбраковка единицы	Удаление (отклонение единицы)
	Создание единицы	Массовая импутация (например, импутация недостающих домашних хозяйств в переписи под одним номером)
		Импутация единиц нижнего уровня в единицу верхнего уровня (например, импутация отсутствующих лиц в ответивших домохозяйствах
		Создание единиц верхнего уровня из единиц нижнего уровня (например, группировка лиц в домашних хозяйствах, вычисление показателей по домашним хозяйствам).
	Объединение единиц	Корректировка недостатков связей (например, техническая проверка связанных записей)
		Сопоставление различных типов единиц (например, размещение домохозяйства с неизвестным адресом в «пустующем» жилище)

Практические решения

Функции часто являются специализированными и могут быть классифицированы в зависимости от конкретных характеристик, таких как типы ошибок (например, очевидные ошибки, систематические ошибки), целей (например, оценки макроуровня) или предстоящие действия (например, интерактивная обработка, импутация). Некоторые функции уже включают в себя два или более этапов процесса в сочетании. Распространенным примером является одновременное применение «проверки» и «отбора», например, определение ошибок с помощью «Филледжи – Хольта» или обнаружение выбросов.

На практике методы, применяемые в производстве, часто не являются методами, представленными в предыдущих разделах. В некоторых случаях из-за вычислительных трудностей реализуемое решение может не полностью отражать первоначально разработанные концептуальные функции и методы. Однако параметризация метода-это задача, которую следует скрупулёзно реализовывать на практике для достижения эффективности и результативности процесса редактирования. Вместо того чтобы распределять параметры редактирования данных по нескольким программам, лучше обеспечить их нахождение в системе метаданных, которая позволяет централизованно использовать их при реализации методов.

Практическим решением снижения риска распространения параметров редактирования данных в различных системах является одновременное выполнение различных функций либо за один раз, либо посредством выполнения последовательности действий. Эти специальные методы верхнего уровня называются методами для комбинации функций. Очень распространенным случаем этого является алгоритм " IF-THEN".

Этот метод объединяет функции всех трех типов функций в одной операции: часть IF предусматривает «проверку» в виде оценки соответствия правилам редактирования (например, условия «ошибки тысячи»), «отбор» заключается в принятии решения об обработке одной или нескольких переменных в соответствие с этим правилами (тех, которые определены в части «THEN»), и «обработка» определяется установкой, в соответствии с которой формируется новое значение.

Другими типичными операциями, относящимися к этому классу, являются анализ выбросов с одновременной проверкой и отбором, а иногда и обработкой, а также парадигма Филледжи-Хольта, которая может включать механизм реализации правил редактирования и алгоритм, необходимый для определения ошибок с минимальными изменениями значений данных.

14. Метаданные, используемые в процессе редактирования данных

Введение

Метаданные можно определить как информацию, необходимую для использования и интерпретации статистических данных (глоссарий Евростата). В этой главе описываются метаданные, необходимые для характеристики и описания процесса редактирования данных.

Во-первых, метаданные необходимы для описания входных и выходных данных процесса редактирования данных, поэтому следующие разделы этой главы посвящены детализации метаданных, сопровождающие входные и выходные результаты. Одни и те же элементы метаданных используются для описания входных и выходных результатов всего процесса редактирования данных, а также входные и выходные результаты, получаемые в ходе каждого этапа процесса.

Во-вторых, чтобы иметь возможность не только описывать, но и управлять процессом редактирования данных, необходимо наличие элементов метаданных для структурирования этапов процесса и процесса в целом. Этапы процесса редактирования характеризуются функциями и методами. Как уже говорилось в главе 2, состав функций определяет, какие действия должны выполняться, а методы - каким образом они выполняются. Кроме того, методы могут быть связаны с правилами.

Функции, методы и правила были подробно описаны в предыдущей главе, которые являются основными элементами редактирования статистических данных. Далее будут представлены некоторые примеры входных и выходных метаданных для трех типов функций (т. е. проверка, отбор и обработка).

Дополнительные объекты метаданных необходимо учитывать в целях характеристики потока процесса, то есть последовательности этапов процесса. Последовательность этапов процесса в процессе управляется с помощью элементов контроля процессов, то есть "набором точек принятия решений, которые влияют на поток/последовательность между этапами процесса, составляющих бизнес-процесс" (модель GSIM).

Сводная таблица, содержащая обзор всех элементов метаданных для процессов редактирования данных с примерами и соответствующими информационными объектами GSIM, представлена в последнем разделе главы

15. Входные метаданные по процессу

Входные метаданные по процессу представляют собой всю информацию, описывающую входные данные процесса редактирования данных, т. е.:

- **Набор данных, являющийся объектом редактирования данных.**
- **Дополнительная информация, необходимая для применения функций, например метаданные, описывающие вспомогательные данные или параметры, необходимые для запуска процесса редактирования данных**

16. Метаданные, описывающие входные данные

Концептуальные и структурные метаданные необходимы для характеристики и описания входного набора данных и вспомогательных данных. Концептуальные и структурные метаданные представляют собой содержание этих данных путем описания измеряемых характеристик (понятий и определений) и их практической реализации (разрезы разработки и структуры данных).

Понятия и определения. Эти метаданные описывают и характеризуют понятия и определения, которые подлежат измерению (например, доход, образование, оборот). Они также определяют объекты этих измерений, которые являются единицами некой совокупности (например, индивидуумы, семьи, предприятия).

Переменные и разрезы разработки. Переменная позволяет объединить понятие с единицей, что приводит к проведению измерений в соответствии с понятиями для каждой единицы (например, доход человека; доход семьи; оборот предприятия).

Переменные могут играть различные роли и эти роли также являются частью описания понятий. Важную роль играет идентификатор единицы. Другими ролями, которые могут быть важны с точки зрения функции редактирования данных, являются: значения/кодов классификатора (с классами, которые могут быть предоставлены центральным сервером классификаторов[1)]; переменная стратификации (определение слоев, для которых некоторые функции редактирования данных выполняются отдельно).

Набор данных по единице, как это обозначено здесь, содержит значения переменных для набора единиц. Для описания этих значений указываются единица измерения и разрезы разработки соответствующих переменных. Для

[1] в случае использование ПО при присвоении значений/кодов классификаторов.

количественных переменных это может быть, например: тысячи евро; неотрицательные вещественные числа. Для категориальных переменных это может быть выражено перечислением кодов категорий и их значений, например 1 - мужской, 2 - женский.

Структура данных. Набор данных по единице - это организованная совокупность значений. Эта организация описывается структурой данных.

Наиболее распространенной структурой данных является «запись». Запись представляет собой набор элементов определенной последовательности, количество которых обычно фиксируется. Им присваиваются серийные номера или идентификационные номера. Элементы «записей» также можно назвать полями. Примерами других структур данных являются: массив, множество, дерево, графики.

Структура «записи» всегда должна содержать хотя бы одну переменную, которую можно использовать в качестве идентификатора единицы. Набор данных может содержать единицы различных типов. Они могут быть иерархически упорядочены, например, люди и домохозяйства. Различные типы единиц измерения могут иметь различные описания записей. Структура данных может также иметь различные атрибуты, описывающие свойства набора данных в целом, такие как этап статистического процесса, к которому он относится, время его создания или совокупность или время, к которому он относится.

17.Вспомогательные/дополнительные данные

Вспомогательные данные включают данных из других источников, за исключением тех, которые редактируются. Вспомогательные данные могут относится к микроуровню - когда вспомогательные данные доступны для всех единиц; или для некоторых из тех единиц, данные по которым редактируются.

Они также могут быть доступны на макроуровне, в случае если вспомогательные данные являются агрегатами, обычно суммами переменных, для тех же самых или коррелирующих с теми, которые редактируются. Разница заключается в том, что в то время как входной набор данных является объектом процесса редактирования статистических данных (т. е. оценивается достоверность и, при необходимости, выполняются определенные действия), вспомогательные данные служат только справочной информацией для одной или нескольких функций, реализующихся в процессе редактирования. При этом сами они не пересматриваются и не изменяются.

Вспомогательные наборы данных на микроуровне - это наборы данных по единице, состоящие из значений переменных, относящихся к единице. В этом смысле они аналогичны набору входных данных. Вспомогательные переменные на микроуровне могут использоваться в функции «проверки» для оценки достоверности данных.

Эта практика предусматривает использование таких переменных при редактировании, функциях счета, обнаружении выбросов. Справочные значения из других источников также могут помочь, например, в обнаружении ошибок «тысячи». Вспомогательные данные на макроуровне могут

использоваться в качестве входных данных для функции «отбора» для оценки достоверности агрегатов при макро-редактировании.

Параметры

Некоторые методы требуют определенных значений для одного или нескольких параметров. В общем случае параметр можно определить как входные данные, используемые для определения параметров конфигурации, которая определяет состав конкретной функции.

Присвоение фиксированных значений параметрам также является частью метаданных, которые необходимо определить до запуска процесса.

Методы импутации в функции «обработки» требуют определения переменных, которые будут использоваться для получения значений импутации. Это могут быть предикторы в параметрических моделях импутации, переменные в функции расстояния для импутации «ближайшего соседа» или переменные, определяющие классы для импутации hot-deck внутри классов.

Отбор значений выбросов или комбинаций значений в функции «проверки» требует определения пороговых значений. Отбор значимых подозрительных единиц для ручного редактирования в функции «отбора» также требует определения пороговых значений.

Определение ошибок, основанная на обобщенной парадигме Филледжи-Хольта, требует определения весов надежности. Корректировка в целях согласованности с жесткими правилами редактирования требует определения этих весов.

Неструктурированные метаданные

Вспомогательные метаданные также могут собираться специалистами по предметной области более или менее неструктурированным образом. Справочные значения для основных переменных могут быть доступны, например, из годовых отчетов предприятий. Кроме того, в интернете может быть доступна информация, например, о текущей деятельности и продуктах компании или о предусмотренной законом прибыли.

Неструктурированные метаданные могут использоваться в интерактивном редактировании. Актуальная информация с веб-сайтов может помочь в редактировании свойств объекта, таких как устаревшие коды классификатора NACE.

18. Метаданные на выходе

Основным результатом процесса является отредактированный набор выходных данных. Метаданные для этого набора данных состоят из описаний концептуальных и структурных метаданных, как описано ранее. Другие метаданные, полученные в процессе редактирования, представляют собой информацию о качестве как для входных, так и для выходных данных. Кроме того, может быть собрана информация о том, как реализуется процесс, которые не имеет прямого отношения к качеству данных (paradata).

Индикаторы качества

В функции «проверки» формируется показатели качества или показатели, которые используются в других функциях «отбора» и «обработки». Они также представляют интерес сами по себе, поскольку эти показатели отражают качество входных данных. В частности, оцененные правила редактирования и оценка единиц.

Матрица неудач. Оценивание по жестким правилам редактирования приводят к матрице булевых значений N × K (число единиц × число правил редактирования).

Эта матрица может быть обобщена несколькими способами.

В частности, можно получить информацию по единице - о количестве несоответствий правилам редактирования либо по видам редактирования - о количестве несоответствий по каждому виду редактирования.

Когда каждый вид редактирования связан с переменными, участвующими в редактировании, также возможно получить разрез по переменным, т.е. количество неуспехов по каждой переменной.

Баллы. Баллы, полученные по единице, являются информацией о качестве единицы и влиянии единиц.

Матрица «неудач» и балльная оценка должны использоваться после окончания каждого этапа процесса, для того, чтобы отследить влияние каждого этапа редактирования данных отдельно на эти индикаторы качества.

Флаги импутации. Флаги импутации обычно указывают на то, была ли переменная импутирована с помощью двоичного индикатора (0/1) или нет. Они могут быть добавлены к выходному набору данных или храниться в отдельном файле.

Они могут быть объединены до уровня единицы, для того чтобы определить, была ли единица импутирована или нет. Флаги импутации служат связующим звеном между входными и выходными данными и могут обеспечивать поддержку качества, расчет коэффициентов импутации, оценку влияния импутации на результаты.

Параданные

Параданные могут формироваться при мониторинге различных действий, которые имели место на этапах процесса. Они могут включать подсчеты этих действий и времени, затраченного на их реализацию.

Параданные могут инициировать пересмотр параметров процесса или изменения в плане процесса для повышения эффективности и результативности процесса.

Входные и выходные метаданные по типам функций

Как уже говорилось ранее, каждый тип функции характеризуется входными и выходными метаданными:

Проверка

Входные метаданные: метаданные, описывающие набор входных данных (например, структуру данных, понятия и определения, переменные и разрезы разработки), а также такие параметры, как допустимые значения или ограничения, наложенные на переменные, определенные в правилах редактирования, или пороговые значения при обнаружении выбросов.

Выходные метаданные: измерение качества в качестве оценки функции проверки (например, уровень ответов по единице и по отдельному показателю, количество единиц, не прошедших контроли, уровень «провалов» правил редактирование).

Отбор

Входные метаданные: критерии отбора (например, парадигма Филледжи - Хольта, детерминированные правила, балльная оценка, порог).

Выходные метаданные: индикаторы, определяющие подмножества единиц и/или переменных набора входных данных, предназначенные для дальнейшей обработки (например, критические единицы для селективного редактирования, некорректная запись).

Обработка

Входные метаданные: метаданные из функций «проверки» и «отбора», позволяют идентифицировать подмножества единиц измерения и переменных для применения функций «обработки», параметры для применения метода обработки (например, предикторы в параметрических моделях импутации, переменные в функции расстояния для импутации ближайшего соседа).

Выходные метаданные: флаги импутации и измерение качества как оценка функций «обработки» (например, уровень импутации и влияние).

Сводная таблица метаданных

В таблице далее приведены в общем виде приведены основные понятия, используемые в процессе редактирования с соответствующими информационными объектами GSIM.

Таблица 5

Понятие метаданных	Примеры	Информационные объекты GSIM
Входные данные по процессу		
Входные данные	Данные обследования, к которым следует применять функции «проверки»,	Входные данные, подлежащие трансформации

	«отбора» или «обработки»	
Дополнительные данные	основа, данные повторных обследований периода t-1, соответствующие административные данные	Входные данные для обеспечения процесса
Параметры	Параметры для обнаружения выбросов, пороговые значения для функций оценки (счетчика)	Входные параметры
Этапы процесса и их последовательность		
Этап процесса	Редактирования по областям, селективное редактирование	Этап процесса
Функция	Проверка допустимости единиц, отбор единиц, подверженных влиянию значимых ошибок, корректировка систематических ошибок	Бизнес-функция
Метод	Правила редактирования для допустимых значений, локализация случайных ошибок, импутация с применением регрессии	Метод в ходе процесса
Контроль процесса	Наличие значимых единиц, наличие подозрительных агрегатов	Контроль процесса
Выходные результаты по процессу		
Выходной набор данных	Данные обследования, преобразованные в ходе «обработки»	Преобразованные выходные данные
Показатели качества и параданные	Количество несоответствий правилам редактирования, флаги импутации	Индикаторы процесса
Концептуальные и структурные метаданные для описания входных, выходных и вспомогательных данных		
Набор данных по единице	Набор данных по обследованию, набор административных данных	Набор данных по единице
Понятия	Доход, оборот	Понятия
Единица	Домохозяйства, индивидуальные предприниматели, предприятия	Единица
Совокупность	Предприятия на 1 января n-го 2019	Совокупность
Показатель	Доходы домашних хозяйств, оборот промышленного предприятия	Показатель

Разрез разработки	Пол (ж/м), доход> 0	Разрез разработки
Структура данных	Запись, массив	Структура данных

19. Модели редактирования

В этой главе описываются элементы, которые могут быть использованы для планирования процессов редактирования статистических данных с учетом элементов плана и ограничивающих факторов, которые определяют модель процесса статистического редактирования данных.

Согласно терминологии модели GSIM, можно представить процесс статистического редактирования данных как "бизнес-процесс", который состоит из "этапов процесса" и "управления процессом", которые могут быть объединены различными способами в соответствии с различными сценариями.

Точнее, процесс статистического редактирования данных можно определить как: определение последовательности и условной логики, существующей между различными этапами процесса. Последовательность этапов процесса в самом процессе управляется элементами управления процессом. Модели статистического редактирования данных процесса, или, сокращенно, модели потока СРД направлены на то, чтобы помочь понять, какие действия выполняются в рамках процесса и каким образом эти действия связаны и контролируются.

20.Элементы моделей процесса

Модели процесса сосредоточены на самой деятельности по редактированию данных. Однако процесс иногда включает в себя действия, которые на самом деле не связаны с редактированием, и результаты которых могут повлиять на выполнение некоторых действий по редактированию данных. Эти действия обычно предусматривают кодирование, объединение, получение новых переменных (синтетических переменных) и получение весов. Результаты затем также проходят через проверку-отбор-обработку. Эти действия не представлены в моделях процессов, с целью их упрощения. Однако "объединение и согласование " кратко описаны в пункте 0.

Этап процесса - это набор конкретных функций, каждая из которых предусматривает реализацию методов, которые выполняются в определенной последовательности для выполнения конкретной цели редактирования. Этапы процесса позволяют различать разнообразные состояния данных и, следовательно, контролировать предыдущие этапы процесса, а также циклы процесса.

Навигация (переход от одного к другому) между этапами процесса осуществляется на основе правил управления процессом. Управление процессом называется обычным, если за этапом процесса следует один и тот же этап процесса при любых обстоятельствах, и необычным, если за этапом

может следовать несколько альтернативных шагов, в зависимости от условий управления процессом.

Разграничение этапов процесса и элементов управления выполняется таким образом, чтобы подчеркнуть соображения в части дизайна (планирования) редактирования данных в целом в соответствии с типом анализируемого процесса.

Типичными примерами таких соображений являются - " сначала обработка ошибок, которые могут быть устранены с высокой надежностью и небольшими затратами", "применение интерактивного редактирования только к единицам со значимыми подозрительными значениями" и "всегда применять макроредактирование".

Первое соображение приводит к этапу процесса "начальное редактирование и импутация", который включает. ряд функций, связанных, например, с обработкой систематических ошибок.

Второе соображение приводит к этапам процесса "интерактивное редактирование и импутация" и "автоматическое редактирование и импутация", которые включают ряд функций, применяемых к различным частям данных.

Третье соображение приводит к проверке, необходимой для завершения процесса редактирования, и может привести к циклу на этапе процесса, когда макро-редактирование завершается неудачей.

Ниже перечислены основные этапы процесса и элементы управления процессом, которые обычно используются для описания процесса редактирования.

21. Этапы процесса

Редактирование по отдельным областям (с точки зрения единиц и переменных). Проверка структурных информационных объектов, определяющих целевую совокупность и переменные: например, верификация и отбор допустимых единиц, классификаторы (например, классификатор ISIC/NACE, правовой статус).

Редактирование систематических ошибок. Этот этап процесса предусматривает работу с явными (очевидными) ошибками, которые легко обнаруживаются и поддаются обработке, и с систематическими ошибками, которые менее заметны, чем предыдущие, но для которых обработка на данном этапе процесса может обеспечить высокий уровень надежности.

Селективное редактирование. Селективное редактирование - это общий подход к обнаружению значимых ошибок. Он основан на идее поиска значимых ошибок с точки зрения основных результатов; при этом производиться тщательная обработка соответствующего подмножества единиц с целью уменьшения затрат на интерактивное редактирование при сохранении желаемого уровня качества оценок.

Интерактивное редактирование. При интерактивном редактировании микроданные проверяются на наличие ошибок и при необходимости корректируются специалистом-редактором на основе экспертного суждения.

Интерактивное редактирование часто следует за селективным редактированием, которое частично позволяет определить ошибки, т. е. осуществить проверку и отбор.

Автоматическое редактирование. Целью автоматического редактирования является обнаружение и обработка ошибок, а также пропущенных значений в файле данных полностью автоматизированным способом, то есть без вмешательства человека.

Макро-редактирование (также известное как редактирование выходных результатов или отбор на макроуровне). Это общий подход к выявлению (отбору) записей в наборе данных, которые могут содержать потенциально значимые ошибки и выбросы, путем анализа агрегатов и/или значений, вычисленных (или экстраполированных) по совокупности.

Согласование переменных. Она заключается в согласовании значений переменных на микроуровне, получаемых из различных источников. Предусматриваются также процедуры, используемые для предсказания (скрытой) целевой переменной с учетом значений наблюдаемых.

Объединение и согласование. Объединение и согласование относятся к обработке микроданных, которые часто необходимы при объединении (согласовании) и взаимоувязывании (достижение соответствия) различных единиц, относящихся к различным источникам входных данных.

Общий сценарий заключается в том, что в связанных наборах данных присутствует много релевантных объектов/единиц, которые могут быть потенциально полезны для следующего этапа процесса получения интересующих исследователей статистических единиц, таких как человек, родство, предприятия и т. д. Этап объединения основан на выявлении всех" связей", которые существуют или допустимы, обеспечивая тем самым основу для последующего вывода (получения) единиц.

Получение структуры [сложной единицы]. Получение и проверка структуры сложной единицы (например, отнесение индивидов к домашним хозяйствам, домашних хозяйств к зданиям).

Следует отметить, что различие между "объединением и согласованием" и "получение сложной структуры единицы" следует из того факта, что эти шаги обычно применяются последовательно: сначала "объединение и согласование", а после "получение сложной структуры единицы".

Чтобы понять разницу, полезно привести пример. Если, согласно некоторым источникам у студента другой адрес, чем у родителей, возможно, потребуется проверить достоверность этой информации, спросив, находится ли этот адрес по месту учебы или нет. Результат этого запроса может быть положительный или отрицательный, при этом он является следствием того, что мы называем "согласование ". Фактическое приписывание жилья или домашнего хозяйства для этого студента-это формирование статистической единицы, которое выполняется только после этого, и производится на этапе процесса "получение структуры сложной единицы".

В дополнение отметим, что ранее представленный этап процесса "объединение и согласование", "получение структуры сложной единицы" и "переменная согласования " относятся к общей совокупности действий,

называемой микро-интеграцией, которая по сути направлена на обработку объединенных данных для того, чтобы обеспечить согласованность и внутреннюю согласованность переменных на микроуровне.

22. Контроль процесса

1) Значимые единицы. Отбор единиц с потенциально влиятельными значениями для интерактивной обработки.

2) Тип переменной (непрерывная, категориальная и т. д.). Отбор переменных для специальной обработки (например, импутации каким-либо подходящим методом, методы редактирования категориальных/непрерывных переменных).

3) Подозрительные агрегаты. Отбор подозрительных агрегатов для обнаружения возможных важных ошибок.

4) Неразрешенные/«запрещенные» микроданные. После повторного применения правил редактирования после этапа обработки некоторые микроданные все еще остаются неразрешенными (некорректными с точки зрения правил редактирования). Это может свидетельствовать о том, что набор правил редактирования для проверки и отбора не был исчерпывающим или что обработка не разрешила все ситуации, связанные с ошибками. В первом случае набор правил редактирования должен быть обновлен, а во втором случае «запрещенные» единицы отбираются для дальнейшей обработки альтернативными методами.

5) Иерархические данные. Проверка того, имеют ли данные иерархическую структуру, то есть существуют ли единицы, которые могут быть сгруппированы в более сложные единицы (например, отдельные лица в домашних хозяйствах, местные единицы в предприятия).

Следует отметить, что этап процесса и контроль процесса могут называться одинаково, при этом формы (последовательность) модели отличаются одна от другой и используются совершенно разные методы.

Например, " автоматическое редактирование" может сильно отличаться от одной ситуации к другой, как с точки зрения методов, так и сложности.

Фактически, в некоторых ситуациях может быть реализован детерминированный подход, основанный на правилах "Если-То", в других случаях- парадигма Филледжи-Хольта.

Тем не менее существуют по крайней мере две основные причины, оправдывающие использование общих названий: (1) экономия усилий на проработку, (2) акцент на сходствах или различиях.

Например, можно подчеркнуть, что ключевое различие между двумя моделями заключается в том, что в одной из них нет необходимости в контроле процесса "значимой ошибки", в то время как в другой такой контроль имеет первостепенное значение.

В таблице ниже перечислены основные этапы процесса и средства контроля процесса статистического редактирования данных.

Таблица 6

Этапы процесса	Функции	Типы функции	Методы
Редак-тирование по областям	Проверка и отбор приемлемых единиц	Проверка, отбор	Если – То
	Проверка, отбор, обработка характеристик данных (код NACE, правовой статус и др.)	Проверка, отбор, обработка	Если – То
Редак-тирование система-тических ошибок	Проверка, отбор, обработка очевидных ошибок	Проверка, отбор, обработка	Если – То
	Проверка систематических ошибок	Проверка	Кластерный анализ, анализ скрытых классов, правила редактирования, графическое редактирование (например, log для 1000-ой ошибки)
	Идентификация единиц, подверженных систематическим ошибкам (значимые единицы)	Отбор	Если – То, кластерный анализ, анализ скрытых классов
	Корректировка систематических ошибок	Обработка	Дедуктивная импутация, импутация на основе модели
Селектив-ное редактиро вание	Идентификация единиц, на которые влияют значимые ошибки	Проверка	Подсчет баллов (оценок)
	Отбор единиц для интерактивной обработки, отбор единиц для неинтерактивной обработки, отбор единиц, не подлежащих обработки	Отбор	Отбор по фиксированному порогу
Интеракти вное редактиро вание	Обработка единиц в критическом множестве	Проверка, отбор, обработка	Повторный контакт, проверка анкет

Автоматическое редактирование	Проверка согласованности данных с точки зрения редактирования	Проверка	Анализ несоответствий правилам редактирования
	Определение показателей, подверженных влиянию ошибок по каждой единице	Отбор	Если – То, парадигма Филледжи-Хольта, метод импутации методом ближайшего соседа (МБС)
	Импутация ошибочных значений	Обработка	Если – То, дедуктивная импутация, неслучайная импутация, случайная импутация, пропорциональное распределение, МБС
	Импутация отсутствующих значений	Обработка	Если – То, дедуктивная импутация, неслучайная импутация, случайная импутация, МБС
Макро редактирование	Проверка и идентификация подозрительных агрегатов и выбросов (значимых единиц)	Проверка, отбор	Анализ выбросов, сравнение агрегатов внутри набора данных, сравнение агрегатов с внешними источниками, сравнение агрегатов с предыдущими результатами

Состояния данных

Этапы процесса, контролируемые средствами управления процессом, содержат входные данные, которые обрабатываются для получения выходных данных. Основными состояниями данных являются:

- **Первичные. Исходный набор данных, который не редактируется (эта категория включает данные, которые могли быть отредактированы поставщиком данных (административные данные) или во время сбора (например, в рамках интернет-опроса или интервьюерами).**
- **Отредактированные. Набор данных после обработки по областям, очевидных и систематических ошибок.**
- **Отредактированные данные после объединения и согласования различных единиц, относящихся к различным входным источникам.**
- **Особо значимые. Набор данных, содержащий потенциально значимые ошибки.**
- **Незначимые. Набор данных без ошибок или содержащий только незначительные ошибки.**

- **Отредактированная [название вышестоящей единицы]-СТ. Набор данных после редактирования структуры (СТ) более высокого уровня. Например, когда высокоуровневой единицей является домашнее хозяйство: Отредактированная [название вышестоящей единицы]-СТ = "отредактированная Д-СТ".**
- **Микроредактирование [название единицы]. Набор данных после редактирования переменных, относящихся к единицам на микроуровне. Например, когда единица домохозяйство: микро-редактирование [название единицы]="микро- редактирование Д".**
- **Финальный набор данных в конце общего процесса редактирования после успешного макро редактирования.**

Элементы планирования

Планирование бизнес-процесса редактирования данных, то есть состав этапов процесса и элементов управления процессом и их комбинирование, определяется конкретными характеристиками входных и выходных данных (называемых "входными и выходными метаданными"), а также ограничивающими факторами.

Входные элементы

Входные метаданные.

Единицы. Тип единиц: предприятия-крупные/малые, физические лица и/или домохозяйства – иерархические единицы, единицы из административных источников, сельскохозяйственные фирмы, макро /микро данные.

Переменные. Типы переменных: числовые, категориальные. Статистические распределения: асимметричные, мультимодальные, с присутствием нулей. Отношения между переменными: правила редактирования.

Обследование. Тип обследования: перепись/выборка, структурные обследования, краткосрочная статистика, панельные опросы, регистр, большие данные.

Характеристика вспомогательной информации. Надежность, своевременность, охват, структурированные /неструктурированные, микро / макро.

Выходные элементы

Тип выходных данных, подлежащих распространению (например, файл микроданных, таблица оценок по предметной области, целевые параметры).

Требования к качеству (например, требуемый уровень точности).

Сдерживающие факторы

Сдерживающие факторы в основном относятся к характеристикам, относящимся к организационным аспектам, которые оказывают сильное влияние на методологический выбор. Наиболее важными сдерживающими факторами являются:

- **Имеющиеся ресурсы (бюджет, человеческие ресурсы, время).**
- **Наличие вспомогательных данных (своевременность, качество, процедура согласования и т. д.).**
- **Цель редактирования с точки зрения полноты и согласованности с учетом внешних источников данных и дальнейшей интеграции данных.**
- **Человеческие компетенции (знания и потенциал).**
- **ИТ (доступные программные и аппаратные средства).**
- **Правовые ограничения.**
- **Политические решения.**

Например, нехватка людей, занимающихся ручной проверкой/контролем наблюдений, может привести к разработке полностью автоматизированной процедуры редактирования данных. Примером политического решения является решение ограничить повторные контакты, чтобы уменьшить нагрузку на респондентов.

С теоретической точки зрения элементы, представленные выше, можно рассматривать как элементы контроля процесса, поскольку они определяют выбор одной модели вместо другой. Фактически, поскольку этап процесса может быть определен с разным уровнем детализации, общий процесс редактирования может рассматриваться как ” этап процесса “на более высоком уровне, и, следовательно, входные и выходные характеристики и ограничивающие факторы могут рассматриваться как” управление/контроль процесса " на этом более высоком уровне.

23. Примеры (сценарии) редактирования данных

Далее представлены примеры моделей редактирования. Этапы процесса представлены в модели прямоугольниками. Состояние данных изображено эллипсами с названиями, связанными с функцией, реализованной на предыдущих этапах процесса.

Сценарий А. Модель. Структурная статистика предприятий (приложение 1)

Структурная статистика предприятий обычно основывается на выборочных, которые могут включать большой набор переменных, как правило количественных. На основе общей модели процесса редактирования и с учетом ключевых элементов, разработана модель для структурной бизнес-статистики.

Сценарий В. Модель. Краткосрочная статистика предприятий (приложение 2)

Краткосрочная статистика предприятий обычно основывается на панельных обследованиях, которые характеризуются небольшим количеством переменных и коротким производственным процессом. Выходные данные представлены в виде индексов и значений вариаций на уровнях агрегации.

Модель в этом сценарии в основном направлена на устранение значимых ошибок по основной целевой переменной для обеспечения точности агрегатов/оценок в короткие сроки. Из-за временных ограничений "автоматическое редактирование" выполняется (например, если микроданные должны быть выпущены/опубликованы) только после завершения интерактивной проверки значимых данных.

Модель представляет собой общую структуру бизнес-обследований. Следует подчеркнуть, что для таких процессов ограничения могут сильно влиять на способ управления потоком. Выбор конкретной стратегии в основном зависит от:

- **Имеющихся ресурсов (например, время, людские и финансовые ресурсы);)**
- **Эффективности автоматического редактирования.**

Например, возможно, нет необходимости возвращаться к селективному редактированию после обнаружения подозрительных агрегатов, например выбросов, которые могут быть обработаны во время взвешивания (не представленного в модели) или путем автоматической импутации.

Кроме того, обнаружение подозрительных агрегатов может обеспечить отбор единиц, связанных с агрегатами, и, следовательно, цикл может приходить непосредственно к интерактивному редактированию. В дополнение, модель явно не указывает на то, применяется ли интерактивное редактирование только к значениям переменных, связанных с обнаружением значимых единиц, или ко всем переменным этих единиц

Сценарий С. Модель. Бизнес –перепись (сплошное наблюдение) (приложение 3)

В случае экономических переписей - из-за большого числа единиц и переменных большое внимание уделяется автоматическим процедурам.

Селективное редактирование выполняется только по тем данным, в которых содержаться подозрительные агрегаты с целью выявления возможного наличия остаточных ошибок (т. е. ошибок, которые не были выявлены на предыдущих этапах процесса редактирования.

Сценарий Д. Статистика домашних хозяйств) (приложение 4)

В случае сплошного обследования предприятий из-за большого числа единиц и переменных большее внимание уделяется автоматическим процедурам.

Селективное редактирование выполняется только в отношении тех данных, которые указывают на подозрительные агрегаты, с целью проверки возможного наличия остаточных ошибок (т. е. ошибок, которые не были выявлены на предыдущих этапах процесса).

Модель для статистики домашних хозяйств в основном зависит от двух конструктивных элементов:

- **тип исследуемых единиц;**
- **тип наблюдаемых переменных.**

Статистика домашних хозяйств может основываться либо на иерархических данных (лица, принадлежащие к домашним хозяйствам) либо на индивидуальных данных. В случае иерархических данных процесс редактирования может быть структурирован по-разному:

- **редактирование переменных по домохозяйству и отдельных переменных выполняются отдельно. В этом случае процесс редактирования состоит из двух последовательных подпроцессов, в рамках которого операции по редактированию данных, выполняемые в рамках последнего процесса, зависят (ограничены) от результатов первого.**
- **переменные, относящиеся к домохозяйству, и отдельные переменные редактируются и импутируются совместно (это выполняется, например, с помощью метода ближайшего соседа/Canceis). В этом случае этапы процесса, относящиеся к структуре домохозяйства, переменным домохозяйства и отдельным переменным, выполняются в рамках отдельного подпроцесса.**

Модель усложняется, если по единицам собираются данные о смешанных типах переменных (категориальные и непрерывные) (например, в случае экономических переменных, таких как доходы, расходы и т.д.) В этом случае может быть выполнено редактирование категориальных и непрерывных переменных:

- **Отдельно: в этом случае процесс редактирования будет включать в себя различные этапы процесса, каждый из которых предусматривает работу с различным типами переменных. В этом случае иерархия между двумя**

подпроцессами должна быть определена, если категориальные и непрерывные переменные связаны друг с другом;

- **Совместно: в этом случае автоматическая обработка категориальных и непрерывных переменных может быть выполнена в рамках отдельного этапа (как это допускается, например, при методе ближайшего соседа/Canceis). Однако, обычно выполняется предварительная идентификации экстремальных значений непрерывных переменных.**

Сценарий E. Статистика на основе интеграции данных (приложение 5)

В последние годы интеграция данных получила значительное развитие. В настоящее время существует множество возможностей, хотя обычно используется и интегрируется ряд административных наборов данных из внешних источников. В других случаях административные данные могут быть интегрированы также с обследованиями. Далее стратегия редактирования описывается в руководстве MEMOBUST (2014) как структурированная. Редактирование выполняется сначала для каждого источника, а затем совместно после этапа объединения и согласования. Сценарий изменится, если один источник также будет содержать данные обследования.

24. Список использованной литературы

1.Generic Statistical Data Editing Model, version 2.0, June 2019, UNECE

2. Generic Statistical Business Process Model, GSBPM, Version 5.1, January 2019

3. Generic Statistical Information Model, GSIM v1.2

4. Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, ISTAT, CBS, SFSO, EUROSTAT

5. MEMOBUST (2014), Handbook on Methodology of Modern Business Statistics, CROS-portal, Eurostat