Оценка качества данных для оценки жизненного цикла в рамках программы ЭДП Центр

Введение

Данное руководство по оценке качества данных основано на широко распространенном подходе (pedigree Matrix approach) (Funtowicz and Ravetz, 1990), адаптированном Weidema и Wesnæs (Weidema and Wesnæs, 1996) для целей оценки жизненного цикла.

Надежность и, следовательно, применимость результатов оценки жизненного цикла зависят от качества исходных данных, обеспечивающих основу для оценки. Поэтому управление качеством данных должно быть неотъемлемой частью оценки жизненного цикла.

Под **данными** следует понимать любую часть информации, которая используется в инвентаризации жизненного цикла.

Качество данных - конкретная характеристика данных, выраженная через информацию о данных (метаданные), такую как:

- информация об их неопределенности (распространение и характер распределения)
- надежность (в зависимости от методов, используемых для измерений, расчетов, предположений и контроля качества данных),
- полнота (количество точек и периодов сбора данных и их репрезентативность для общей совокупности),
- давность (год исходного измерения),
- географическая область, для которой данные являются репрезентативными,
- технология процесса или технологический уровень, для которого данные являются репрезентативными.

Формальное управление качеством данных начинается с определения стратегии сбора данных как части цели и охвата оценки жизненного цикла.

Во время сбора данных качество данных документируется для каждого набора данных. Качество отдельных наборов данных может впоследствии быть связано с целями исследования через ряд показателей качества данных, которые определяют качество данных по отношению к способу их использования в исследовании.

Наконец, показатели качества данных могут быть интерпретированы как дополнительная неопределенность отдельных данных. Такие дополнительные неопределенности могут быть включены при расчете неопределенности общего результата инвентаризации жизненного цикла и, таким образом, могут быть использованы для придания этому результату более квалифицированного выражения.

Оценка качества данных проводится согласно “матрице происхождения” (“pedigree matrix”), которая дает полуколичественное указание на:

1. надежность (reliability) данных, включая оценку методов сэмплирования и процедур проверки.

2. полнота данных, включая статистическую репрезентативность данных, количество измерений в выборке и периоды времени для сбора данных.
3. временные, географические и другие технологические корреляции (между данными и целями качества данных).

Эти показатели качества данных могут впоследствии использоваться:

- для пересмотра стратегии сбора данных с целью улучшения качества собранных данных;
- в сочетании с оценками неопределенности для более точной оценки надежности результата.

Цели качества данных

Цели качества данных определяют в общих чертах желаемые характеристики данных, необходимые для исследования.

Цели качества данных могут выражать:

1. относительную или абсолютную важность различных аспектов качества данных (например, проверка, давность данных)
2. желаемый географический и технологический уровень
3. важность использования данных сопоставимого качества для различных альтернатив
4. способы преодолевать с пробелы в данных и отсутствие репрезентативности (например, с помощью оценок наихудшего случая или с помощью оценочных средних значений)
5. способы преодолевать отсутствующую информацию о неопределенности.

Одна и та же цель качества данных может привести к совершенно разным требованиям к качеству данных для разных этапов жизненного цикла (этапы, которые оказывают большое влияние на результат, должны быть охвачены данными лучшего качества, чем этапы, которые оказывают меньшее влияние).

Таким образом, цели качества данных не определяют никаких определенных критериев для отдельных наборов данных, а являются общими спецификациями предпочтений и процедур, которые следует применять при сборе данных и оценке качества данных.

Конкретные используемые цели качества данных будут зависеть от целей и области применения исследования.

Например:

- исследование, ориентированное на конкретное предприятие и предназначенное для долгосрочного стратегического использования, будет касаться других географических и технологических уровней, нежели публичное исследование, сравнивающее существующие на рынке продукты
- исследование, посвященное конкретному предприятию, будет уделять больше внимания получению последних актуальных данных, нежели независимо

проверенных данных, в то время как публичное исследование может поменять этот приоритет

- использование данных сопоставимого качества будет иметь наибольшее значение для сравнительного исследования
- сравнительное исследование будет опираться на худшие оценки пробелов в данных, в то время как стратегическое исследование может просто сообщать о пробелах в данных как таковых.

Сбор данных

Качество данных должно быть задокументировано в рамках деятельности по сбору данных. Рекомендуется сообщать следующие характеристики для всех собранных и используемых данных:

6. метод получения (метод измерения для данных, полученных в результате прямых замеров, метод расчета для усредненных или иным образом рассчитанных данных, предположения, сделанные для оценочных данных)
7. методы, используемые для проверки/верификации данных
8. количество точек сбора и периодов сбора данных, их репрезентативность для общей численности населения;
9. давность данных, выраженная как год первоначального измерения
10. географическая область, для которой данные являются репрезентативными
11. технологический процесс или технологический уровень, для которого данные являются репрезентативными.

Когда качество данных неизвестно, это следует указать.

Показатели качества данных

Качество отдельных данных или наборов данных может быть впоследствии связано с целями качества данных через ряд показателей качества данных, которые определяют качество данных в отношении способа их использования в исследовании.

Индикаторы качества данных могут использоваться для оценки согласованности фактического качества собранных данных в отношении конкретных целей качества данных.

Показатели качества данных - это полуколичественные показатели, прикрепленные к набору данных, представляющие качество данных. Следующие показатели качества данных необходимы и достаточны для описания тех аспектов качества данных, которые влияют на надежность результата (см. также Таблицу 1):

1. **Надежность** (reliability) - показатель надежности относится к источникам, методам получения и процедурам проверки, используемым для получения данных. Показатель не зависит от целей качества данных: решение, принятое в соответствии с определением цели, не изменяет надежность данных, и оценка была бы идентичной, если бы данные использовались в другом исследовании.
2. **Полнота** (completeness) - показатель полноты относится к статистическим свойствам данных: насколько репрезентативна выборка, включает ли выборка

достаточное количество данных и является ли период достаточным для выравнивания нормальных колебаний. Что касается “показателя надежности”, “показатель полноты” относится к свойствам самой выборки и не зависит от целей качества данных конкретного исследования.

Следующие три показателя относятся к связи между данными и целями качества данных, касающимися технологии или условий производства в широком смысле

3. **Временная репрезентативность** – “временной показатель” представляет временную корреляцию между годом исследования (как указано в целях качества данных) и годом полученных данных. Поскольку в некоторых отраслях технологии развиваются очень быстро, десятилетняя разница между годом исследования и годом данных может привести к полному изменению выбросов или эффективности производства. Следовательно, этот показатель тесно связан с целями качества данных в соответствии с определением цели.
4. **Географическая репрезентативность** – “географический показатель” иллюстрирует географическую корреляцию между определенной областью (как указано в целях качества данных) и полученными данными. Методы и условия производства и технологии могут сильно отличаться в разных странах. Например, различная структура источников генерации электроэнергии, разные способы и технологии ведения сельского хозяйства и т.п.
5. **Технологическая репрезентативность** – “технологический индикатор” касается всех других аспектов корреляции, кроме временных и географических соображений. Хотя данные могут быть желаемой давности и репрезентативны для желаемой географической области, они могут не быть репрезентативными для конкретных предприятий, процессов или материалов, которые изучаются. Таким образом, может возникнуть необходимость использовать данные о родственных процессах или материалах, которые в некоторых случаях могут рассматриваться как предпочтительные по сравнению, например, с более старыми данными или данными из другой географической области. Показатель “технологическая репрезентативность” относится только к различиям, возникающим, несмотря на то, что имеющиеся данные подходят по давности и относятся к одной и той же географической области. Например, если нужны данные по производству тракторов в Дании за 1992 год, а у нас есть данные по производству тракторов в Великобритании за 1976 год, то все различия в технологиях можно отнести к проблемам во временной и географической корреляции. Однако, если, кроме того, данные Великобритании относятся к производству автомобилей, а не к производству тракторов, это будет отражено в показателе “дальнейшая технологическая корреляция”.

Матрица в Таблице 1 называется матрицей происхождения (pedigree matrix), поскольку показатели качества данных относятся к истории или происхождению данных, как генеалогическое древо отражает родословную человека. Цель матрицы происхождения — предоставить инструмент управления качеством данных, который позволяет легко исследовать качество данных, указывать на возможности улучшения качества данных и отслеживать источники неопределенности.

Баллы в матрице происхождения являются “полуколичественными”. Они служат только в качестве идентификационных номеров и не должны рассматриваться как представляющие определенное “количество” качества данных. Поэтому следует избегать любых попыток агрегировать числа.

Таблица 1. Матрица происхождения для 5 показателей качества данных.

Балл показателя	**1**	**2**	**3**	**4**	**5**
Надежность	Верифицированные данные[1], основанные на измерениях[2]	Верифицированные данные частично основанные на предположениях или не верифицированные данные, основанные на измерениях	Не верифицированные данные, частично основанные на предположениях	Квалифицированная оценка (например, от отраслевого эксперта)	Неквалифицированная оценка
Полнота	Репрезентативные данные из достаточной выборки (производств или других источников) за достаточный период времени для выравнивания нормальных колебаний	Репрезентативные данные из меньшего количества источников, но за достаточный период времени для выравнивания нормальных колебаний	Репрезентативные данные из достаточной выборки (производств или других источников) за более короткий период времени	Репрезентативные данные, но из меньшего числа источников и за более короткие периоды времени или неполные данные из достаточного числа источников и за достаточный период времени	Репрезентативность не установлена или данные неполны и/или из меньшего числа источников или за более короткие периоды времени

[1]Верификация может осуществляться несколькими способами, например, путем проверки на месте, путем пересчета, с помощью баланса масс или перекрестных проверок с другими источниками

[2]Включает расчетные данные (например, выбросы, рассчитанные из входов в процесс), когда основой для расчета являются измерения (например, измеренные входы). Если расчет частично основан на предположениях, должен присваиваться балл 2 или 3.

Временная репрезентативность	Разница между годом сбора данных и исследованием менее 3 лет	Разница между годом сбора данных и исследованием менее 6 лет	Разница между годом сбора данных и исследованием менее 10 лет	Разница между годом сбора данных и исследованием менее 15 лет	Давность данных неизвестна или разница между годом сбора данных и исследованием более 15 лет
Географическая репрезентативность	Данные по географической области, где расположена исследуемая продукционная система	Усредненные данные по более широкой географической области, которая включает в себя ту, где расположена исследуемая продукционная система	Данные по географической области со схожими условиями производства	Данные по географической области с отчасти схожими условиями производства	Данные из неизвестной географической области или территории со значительно отличающимися условиями производства
Технологическая репрезентативность	Данные по предприятиям, процессам и материалам, составляющим исследуемую продукционную систему	Данные по исследуемым процессам и материалам, но с предприятий, не включенных в исследуемую продукционную систему	Данные по исследуемым процессам и материалам, но по другим технологиям	Данные по родственным процессам или материалам, но по одной и той же технологии	Данные по родственным процессам или материалам, но по отличающимся технологиям

Таблица 2. Пример показателей качества данных для потока производства полиэтилена для пластиковых пакетов.

Показатель качества данных	**Балл**	**Пояснение**
Надежность	1	Данные собраны на производственной площадке и являются результатом прямых замеров
Полнота	2	Данные собраны на одной из 5 производственных площадок полиэтилена
Временная репрезентативность	1	Данные собраны за прошлый календарный года производства полиэтилена
Географическая репрезентативность	3	Данные получены на территории со схожими условиями производства
Технологическая репрезентативность	1	Полиэтилен производится на предприятии, которое является частью продукционной системы, технологии производства одинаковы на всех производственных площадках

Таблица 3. Индексы качества данных для процессов производства полиэтиленовых пакетов

Данные	**Индекс качества данных**
Производство 1 кг полиэтилена	(1, 1, 1, 1, 1)
Производство 1 кВт*ч электроэнергии	(2, 2, 1, 2, 4)
Производство этилена из ШФЛУ	(1, 2, 1, 3, 1)
Производство красителя белого	(5, 5, 5, 5, 4)
Производство антислеживающего агента	(4, 5, 3, 4, 5)
Энергия для экструзии 1 кг полиэтилена	(1, 2, 1, 3, 1)
Энергия для производства пакетов	(1, 2, 1, 3, 1)
Энергия для производства полиэтилена	(2, 1, 1, 2, 4)

Использование показателей качества данных для улучшения качества данных

Перечисление показателей качества для всех данных может углубить понимание типичных проблем с качеством данных конкретного исследования. Таблица 3 показывает такой перечень для производства полиэтиленовых пакетов.

Исходя из такого перечня можно увидеть основные проблемы или недостатки данных, например, какой параметр может быть в первую очередь улучшен для повышения общего качества данных.

Важно: если некоторые наборы данных необходимо обобщить, итоговый индекс должен содержать наибольшие баллы из каждого отдельного индекса.

Например, если мы хотим обобщить данные по добавкам в полиэтиленовые пакеты (краситель и антислеживающий агент), то результирующий индекс будет (5, 5, 5, 5, 5), а если обобщение будет происходить по данным для энергии, то результирующий индекс составит (2, 2, 1, 3, 4).

Если бы сообщались только такие обобщенные значения, это было бы не очень информативно, поскольку это сообщает читателю только о том, что где-то в исследовании возникла проблема с качеством данных, но не о том, где конкретно есть проблема с качеством данных и насколько серьезна эта проблема. Таким образом, при использовании по отдельности индексы качества данных не дают адекватной информации о качестве результата. Чтобы преодолеть это ограничение, показатели качества данных должны рассматриваться в сочетании с информацией о неопределенности данных.

Изменение показателей качества данных в сочетании с неопределенностью

Управление неопределенностью помогает исследователю глубже анализировать результаты каждого этапа жизненного цикла.

Если для исследователя доступны показатели неопределенности доступных данных, рекомендуется учитывать их наряду с показателями качества данных.

Можно выделить два вида неопределенности данных:

1. “базовая” неопределенность, которая связана со всеми выборочными данными (обычно это ошибки измерения и нормальные колебания измеряемой переменной),
2. “Дополнительная неопределенность”, которая связана с неоптимальным качеством данных, что отражается в индексе, отличном от (1, 1, 1, 1, 1).

Часто даже базовая неопределенность неизвестна. Если доступна выборка результатов измерений, можно рассчитать базовую неопределенность данных, но чаще всего доступно только одно значение.

Показатели качества данных относятся только к качеству данных с их базовой неопределенностью.

Это означает, что если конкретный аспект качества данных учитывается при описании неопределенности данных, то данные происхождения данных должны быть изменены соответствующим образом. Чтобы показать, что новый индекс относится не к исходным данным, а к измененным данным, для показателя используется “m” вместо исходного показателя качества данных.

Пример (таблица 4):

Таблица 4.

Показатель качества данных	**Балл**	**Неопределенность (коэффициент вариации, CV)**
Базовая неопределенность		5.5%
Надежность	2	2.5%
Полнота	1	-
Временная репрезентативность	1	-
Географическая репрезентативность	2	17%
Технологическая репрезентативность	4	33%
Модифицированная неопределенность ($CV_{\text{итог}}$)		38%

Предположим, что есть некоторое количество полиэтилена (100 кг), которое требуется для производства пакетов. Базовая неопределенность, выраженная коэффициентом вариации, составляет 5.5%. Данные основаны на колебаниях в количестве потребляемого полиэтилена за 5 лет.

Коэффициент вариации отражает степень изменчивости по отношению к среднему показателю выборки:

$$CV = \frac{\sigma}{\mu} \times 100\%,$$

$\sigma -$ стандартное отклонение,

$\mu -$ среднее значение.

По каждому показателю качества данных также получены данные о неопределенности. В таком случае итоговая неопределенность, выраженная как коэффициент вариации, будет равна квадратному корню из суммы квадратов отдельных коэффициентов вариации.

$$CV_{\text{итог}} = \sqrt{CV_1^2 + CV_2^2 + CV_3^2 + \ldots + CV_n^2},$$

то есть выражение неопределенности результата вычисляется как стандартное отклонение суммы этих неопределенностей.

Как видно из приведенного выше примера, влияние высокого балла на показатель качества данных может проявляться как увеличением неопределенности, так и изменением среднего значения. Влияние отдельных показателей качества данных является аддитивным, поскольку показатели определяются как независимые.

В некоторых случаях можно рассчитать дополнительную неопределенность. В других случаях неопределенность необходимо оценить. В приведенных выше примерах неопределенность оценена на основе знаний о рассматриваемых процессах. Другим вариантом было бы

использование неопределенностей по умолчанию, сгенерированных из других аналогичных наборов данных, где неопределенности могут быть рассчитаны.

Из примеров видно, что модифицированная неопределенность в основном определяется самым большим значением неопределенности для отдельных показателей качества данных. Таким образом, если источник самой большой неопределенности очевиден (как, например, в приведенной выше оценке (CV 5,5%) (2, 1, 1, 2, 4), где очевидно, что балл 4 для технологической репрезентативности будет вносить основной вклад в модифицированную неопределенность), можно использовать упрощенную процедуру, где оценка этого вклада в неопределенность (CV 33% в Таблице 4) просто используется как оценка общей модифицированной неопределенности по этому набору данных.

Источник наибольшей неопределенности может быть настолько очевиден, что его можно определить без использования индекса качества данных, и тогда эту часть процедуры можно опустить.

В таблице 5 представлены сводные результаты аналогичной работы по каждому набору данных:

Таблица 5. Модифицированная неопределенность (в виде коэффициентов вариации для всех наборов данных):

Данные	**Базовая неопределенность, CV**	**Индекс качества данных**	**Модифицированные коэффициент вариации ($CV_{\text{итог}}$) и среднее (μ)**
Производство 1 кг полиэтилена	1%	(1, 1, 1, 1, 1)	1%
Производство 1 кВт*ч электроэнергии	19%	(2, 2, 1, 2, 4)	25%
Производство этилена из ШФЛУ	8%	(1, 2, 1, 3, 1)	12%
Производство красителя белого	59%	(5, 5, 5, 5, 4)	62% (и μ на 20% выше)
Производство антислеживающего агента	31%	(4, 5, 3, 4, 5)	59%
Энергия для экструзии 1 кг полиэтилена	-	(1, 2, 1, 3, 1)	17%
Энергия для производства пакетов	10%	(1, 2, 1, 3, 1)	20% (и μ на 10% ниже)
Энергия для производства полиэтилена	5.5%	(2, 1, 1, 2, 4)	38%

Если количество накопленных данных и экспертный опыт позволяют, то каждому баллу показателя качества данных можно сопоставить наиболее характерную дополнительную

неопределенность. Это даст возможность проставлять значения CV по умолчанию, на основе данных из таблицы. Эти оценки могут быть как расчетными, так и основанными на экспертных суждениях.

Важно понимать, что такие таблицы составляются для каждой конкретной отрасли или технологии производства, поэтому их не следует использовать в качестве значений по умолчанию для любого другого исследования, пока не будет накоплено больше опыта.

Таблица 6. Матрица для накопления типичных дополнительных неопределенностей (в виде коэффициентов вариации, если не указано иное) для отдельного процесса.

Балл показателя	**1**	**2**	**3**	**4**	**5**
Надежность	0%	3-10%	17-25%	30%	95%
Полнота	0%	0-10%	?	?	25%
Временная репрезентативность	0%	10%	20%	30%	40%
Географическая репрезентативность	0%	5-17%	10-25%	50%	50%
Технологическая репрезентативность	0%	16%	25%	33%	50%

Способ использования Таблицы 6 для некоторых процессов одного типа (например, энергетические процессы) может выглядеть так: мы получили индекс качества данных для производства и поставки электроэнергии в некоторой отрасли (1, 2, 1, 2, 1).

Мы видим, что если качество наших данных не соответствует наилучшему баллу, то неопределенность увеличивается. Например, балл 5 по показателю надежности дает очень значительную дополнительную неопределенность (95%), в то время как такой же балл по показателю полноты - 25%. Для нашего исследования это может значить, что в первую очередь необходимо работать над улучшением показателей надежности, в то время как полнота

Согласно нашей таблице, дополнительная неопределенность для нашего набора данных составит:

$$CV_{\text{дополнительная}} = 0 + (\text{от } 3 \text{ до } 10) + 0 + 10 + 0 = \text{от } 23 \text{ до } 33\%$$

Резюме

Таким образом, методика анализа качества данных, предлагаемая нашей Программой, предполагает следующий процесс:

12. Сбор исходных данных
13. Присвоение показателей качества данных для каждого набора данных и формирование матрицы происхождения
14. Работа с матрицей происхождения данных: качественное описание результатов и обсуждение ключевых проблем качества данных
15. Расчет или оценка неопределенности:
    a. базовая неопределенность: определяется для каждого набора данных исходя из вариаций в исходных измерениях (например, стандартные отклонения энергопотребления для конкретных процессов).
    b. дополнительная неопределенность: связана со сниженным качеством данных (например, устаревшие данные, слабая географическая репрезентативность). Эта неопределенность добавляется в виде коэффициентов вариации (CV) для каждого показателя качества данных, как описано в Таблице 5.
16. Расчет итоговой неопределенности:

$$CV_{\text{итог}} = \sqrt{CV^2_{\text{базовая}} + CV^2_{\text{дополнительная}}}$$

17. Построение распределения. Могут использоваться различные методы моделирования и симуляции для построения распределений. В практике ОЖЦ чаще всего применяется симуляция Монте-Карло. Каждое отдельное значение учитывает вариации, описанные коэффициентами вариации. Эти вариации используются для многократных симуляций, чтобы получить распределение возможных значений по конкретному процессу.

Шаги 1-3 строго обязательны для отчетов о проведении ОЖЦ для разработки ЭДП. Остальные шаги рекомендованы, но могут быть пропущены, если исследователь не имеет возможности рассчитать неопределенность или не располагает экспертными оценками.

Источники

1. Weidema, P and Wesnaes, M.S. (1996). Data quality management for life cycle inventories-an example of using data quality indicators. J. Cleaner Prod. Vol. 4, No. 3-1, pp. 167-174.
2. Funtowicz, S.O. and Ravetz, JR. (1990). Uncertainty and quality in science for policy. Kluwer Academic Publishers, Dordrecht.