УДК 338.9
ББК 65.9(2Р)30-2
П 781

П 781 **Проблемы и перспективы модернизации российской экономики** / отв. ред. А.В. Алексеев, Л.К. Казанцева. – Новосибирск : ИЭОПП СО РАН, 2014. – 328 с.

ISBN 978-5-89665-272-4

В сборнике опубликованы статьи сотрудников Института экономики и организации промышленного производства СО РАН, содержащие результаты исследований, выполненные по Программе IX.84.1. Экономика как вероятностная система: статистические и теоретические исследования, прикладные выводы.

Рассмотрены народнохозяйственные и отраслевые особенности технологического перевооружения обрабатывающей и добывающей промышленности, изучен международный опыт. Проанализированы институциональные факторы развития технологической системы, а также экологические проблемы и их влияние на общественное здоровье в регионах РФ.

Сборник представляет интерес для научных работников, занимающихся анализом и моделированием экономических процессов, а также для преподавателей, аспирантов и студентов экономических вузов.

УДК 338.9
ББК 65.9(2Р)30-2

ISBN 978-5-89665-272-4

*Э. А. Сатанова*

# К ВОПРОСУ О ВЛИЯНИИ ФАКТОРА НЕОПРЕДЕЛЁННОСТИ В ИСХОДНОЙ ИНФОРМАЦИИ БАЛАНСОВ ПРОИЗВОДСТВЕННЫХ МОЩНОСТЕЙ НА ТЕНДЕНЦИИ ПОВЕДЕНИЯ ПАРАМЕТРОВ ВЫХОДНЫХ ВЕЛИЧИН

Как известно, любые экономические исследования начинаются со сбора исходной информации. Выраженные в числовом виде результаты наблюдений и экспериментов имеют большое значение, так как от их качества, в конечном счёте, зависит итог исследований, а обработка этих данных приводит к теоретическому осмысливанию результатов наблюдений, и к конечной цели – установлению законов, дающих возможность прогнозировать или анализировать поведение явлений [1].

Практически каждое конкретное измерение данных, получаемое в процессе наблюдения, дает, как правило, приближенное значение величины явления. Оно в той или иной мере отличается от истинного значения.

Из всего многообразия рассматриваемых в теории ошибок наибольшую опасность представляют *систематические.*

Если *случайные* ошибки, возникающие под действием случайных факторов, на конечном результате обычно не отражаются, так как взаимопогашаются при сводной обработке результатов наблюдения, то *систематические ошибки,* как правило, имеют одинаковую тенденцию либо к уменьшению, либо к увеличению значения показателя признака. Они представляют большую опасность, так как в значительной мере искажают результаты наблюдений.

В появлении ошибок, *зависящих от стадии возникновения,* превалирующую роль играет *человеческий фактор* (*неточности при записи* данных и их вводе в вычислительную технику, *потеря* части данных из-за несоблюдения технологии хранения информационных баз, *волюнтаризм* при пользовании отдельными показателями), а также искажение данных при передаче через линии связи и т. д.

Причины возникновения ошибок при подготовке информации для экономических исследований можно квалифицировать следующим образом:

*ошибки измерения* – связаны с погрешностями, которые возникают при однократном статистическом наблюдении явлений и процессов общественной жизни;

*ошибки репрезентативности* – возникают в ходе дискретных наблюдений и связаны с тем, что сама выборка может быть не репрезентативной, и результаты, полученные на её основе, не могут распространяться на всю совокупность;

*преднамеренные ошибки* – возникают из-за сознательного искажения данных с разными целями;

*непреднамеренные ошибки,* как правило, носят случайный характер. И их причиной также является человеческий фактор (низкая квалификация, невнимательность, небрежность).

Экономическая информатика даёт определение *достоверности статистической информации* как свойства последней отражать реально существующие объекты с необходимой точностью. Достоверность информации измеряется доверительной вероятностью необходимой точности, т.е. вероятностью того, что отображаемое информацией значение параметра не отличается от истинного значения этого параметра в пределах необходимой точности.

Контроль достоверности статистической информации осуществляется на всех этапах проведения наблюдений, начиная со сбора первичной информации и до этапа получения итогов [2, 3].

Обычно перед работой со статистическими данными проводят три вида контроля: арифметический, синтаксический и логический.

При *арифметическом контроле* полученные итоги сравниваются с предварительно подсчитанными контрольными суммами. Часто арифметический контроль основывается на зависимости одного показателя от двух или нескольких других.

При *синтаксическом контроле* проводится проверка правильности структуры документа, наличия необходимых и обязательных реквизитов, полноты заполнения строк формуляров в соответствии с установленными правилами.

При *логическом контроле* проверяется правильность записи кодов, соответствие их наименованиям и значениям показателей, а также выполняется проверка необходимых взаимосвязей между

показателями, сопоставляются ответы на различные вопросы и выявляются несовместимые сочетания [3].

Проводимый нами в течение многих лет цикл работ, связанных с рассмотрением разных аспектов поведения производственных мощностей (ПМ) – как в течение года, так и в динамике – привёл нас к необходимости создания динамических Баз и Банка данных (БД). Исходной информацией для них послужили статистические данные ежегодных балансов производственных мощностей (БПМ) в натуральном выражении в разрезе нескольких сотен позиций, разрабатываемые Росстатом. К настоящему времени в нашем распоряжении имеются БПМ за 1997–2012 гг.

Работа с отдельными годовыми балансами ПМ в натуральном выражении в ряде случаев оказалась недостаточно продуктивной. Построение динамических рядов на основе погодовых данных с большой вероятностью обнаруживает наличие ошибок. Найти их в огромных массивах информации довольно сложно. Именно поэтому было решено создать электронные Банк и Базы данных за 1997–20XX[1] гг. Данные создаваемых БД хранятся в виде таблиц EXCEL.

***Структура базы данных.*** Анализ годового баланса ПМ показал, что его можно представить как сложную структуру.

Из разных *типов логических моделей* управления данными (иерархической, сетевой, реляционной) был, в силу специфики информации и решаемых задач, выбран смешанный (или так называемый – *гибридный*) тип. Он включает в себя некоторые *свойства* всех трёх типов и подчиняется следующим требованиям.

- Данные создаваемой БД, как и в реляционных моделях, хранятся в виде *таблиц EXCEL*.
- Каждая строка таблицы – *запись* – содержит информацию, относящуюся только к одному *объекту*. Объектом является информация, относящаяся к одной отдельно взятой позиции баланса ПМ.
- Столбец таблицы содержит однотипную для всех записей информацию и называется *полем*. (Например, поле, содержащее информацию о ПМ на начало года, или поле, содержащее наименования позиций в отраслях.)
- БД должны быть *структурированы*.

[1] 20XX означает, что конечный год информации в БД – текущий. Вначале БД оканчивалась 2005 г., затем 2008, сейчас 2012.

При определении *структуры* данных в базе выделяют следующие основные понятия.

• *Класс объектов* – совокупность объектов, обладающих одинаковым набором свойств. (В нашем случае классом объектов является совокупность позиций крупных отраслей промышленности, относящихся к каждому конкретному году.)

• *Свойство (атрибут)* – определенная часть информации о некотором объекте. Хранится в виде столбца таблицы (например, *Наименование позиции* – это свойство объекта *Номенклатура БД*). *Атрибуты записей* – параметры балансов ПМ – находятся в столбцах электронной таблицы.

• *Связь* (*отношение*) – способ, которым связана информация о разных объектах.

Существуют три *типа связей* между объектами:

(1) один к одному (1 : 1);

(2) один ко многим (1 : ∞);

(3) много ко многим (∞ : ∞).

• *Ключевое поле* – это поле, которое содержит уникальные (т.е. неповторяющиеся) для каждой записи данные. (В создаваемых БД – ключи составные – *Присвоенный(-е) номер(а) позиции* и *Наименование позиции*.)

Отбор записей производится по заданному критерию с использованием операторов логических операций [*или (or), если (if), и (and)*], либо их сочетаниями, либо фильтрацией по признаку – в зависимости от задачи.

*Структурированные типы данных* предназначены для задания сложных структур данных.

Совокупность балансов (База данных) является ещё более сложной структурой (системой массивов), для каждого из которых ключом является номер года [4, 5, 6].

Такая организация данных оказалась очень удобной при создании и работе с БД, контроле и обработке информации, при её анализе. Как видим, каждый годовой баланс ПМ в разработанной структуре представим в виде таблицы, в которой по строкам расположены записи (информация о мощности), по столбцам их ключи и атрибуты (свойства данной мощности).

***Согласование данных, их очистка, создание БД.*** Чтобы избежать при расчётах и анализе возможности двойного счёта, мы исключили дублирующие друг друга позиции. В одних информа-

ция по одной и той же позиции дана в альтернативных единицах измерения, в других – информация приведена как в агрегированном, так и в дезагрегированном виде. Кроме самой агрегированной позиции присутствуют данные по нескольким её составляющим, примерно или точно представляющим в сумме одну агрегированную. Важно отметить, что при этом некоторая часть информации может быть утеряна. В качестве примера приведём данные по одной из позиций 2008 г.

Так, вместо исключённой позиции «прокат чёрных металлов готовый, включая заготовку на экспорт», были оставлены две её составляющие – прокат сортовой и прокат листовой (табл. 1).

*Таблица 1*

**Пример внесения ошибки при дезагрегировании**

| Показатель | ПМ на начало 2008 г. |
|---|---|
| Прокат чёрных металлов готовый, включая заготовку на экспорт, тыс. т | 65648,5 |
| Прокат сортовой, тыс. т | 30895,5 |
| Прокат листовой, тыс. т | 24991,0 |
| Доля составляющих в агрегированной позиции, % | 85,1 |

Если судить по названию, то данные по отдельной позиции «заготовка на экспорт», как и сама позиция, в балансе не представлены. Как видно из табл. 1, доля составляющих в агрегированной позиции составляет 85%. Тем самым в информацию по балансу *внесена ошибка*.

Можно представить, что в статистической информации БПМ, таких ошибок достаточно много, так как при составлении балансов ПМ все позиции тем или иным образом агрегированы. Однако оценить величину внесенной при этом ошибки при суммировании, неполноте информации, той или иной методике счёта, мы не можем в связи с тем, что у нас отсутствуют для этого исходные наблюдения.

В принципе, зная количество слагаемых, можно вычислить статистическую оценку ошибки суммы [1]. Например, при сложении 50 слагаемых, заданных с точностью до 0,005, можно сказать, даже не зная закона распределения, что с вероятностью $> 0{,}99$ «предельная» суммарная ошибка $\approx 0{,}2$, т.е. в сумме не надёжны даже десятые доли. Поэтому, в силу перечисленных выше причин, просто отметим, что в балансы ПМ внесены *систематические* ошибки.

Для сохранения структуры БД, исключённые позиции в таблицах были представлены пустой записью, имеющей только значения ключей, что приводит к различию в количестве реально заполненных информацией строк балансовой таблицы.

Создаваемое таким образом в ряде случаев некоторое неудобство, достаточно легко преодолевается как программным способом, так и при проведении экономического анализа, в то время как нарушение структуры в каждом отдельном году порождает массу проблем и создаёт огромный фронт для появления ошибок, связанных со случайным сдвигом информации, отследить которые очень сложно, а главное, вероятность не нахождения подобных ошибок достаточно велика.

За период 1997–2012 гг. дважды произошли крупные изменения при разработке балансов Росстатом (в 2002 г. и 2010 г.), что привело к существенным трудностям при согласовании данных. В 2002 г. Росстат частично изменил номенклатуру БПМ. Начиная примерно с 2010 г. проводится методическая реформа, состоящая в переходе на новый классификатор продукции ОКВЭД. Он вывел из номенклатуры большое количество устаревших морально и физически ПМ и ввёл новые современные мощности.

И если при этом атрибуты записи во всех годовых балансах ПМ практически оставались неизменными (кроме некоторого расширения количества составляющих атрибутов увеличения и уменьшения мощностей в 2002 г.), то в отношении номенклатур этого сказать нельзя. Так, в ряде позиций в 2002 г. массово изменились наименования и частично наполнение позиций.

Так, до 2009 г. балансы, разрабатываемые с использованием классификатора ОКП, структурно представляли собой матрицы размерности ($n$, $m$). Здесь $n = 1,..., k_i$ ($n$ – число крупных отраслей промышленности (класс объектов); $k_i$ – число позиций, входящих в отрасль; $i$ – порядковый номер отрасли); $m$ – число столбцов матрицы, в каждом из которых содержатся однотипные показатели данных по каждой позиции (атрибуты).

Начиная же с 2010 г. структура балансов производственных мощностей в натуральном выражении, разрабатываемых с использованием классификатора ОКВЭД, была представлена как структура по видам экономической деятельности (ВЭД), т.е. класс объектов изменился. Кроме того, изменилась и сама номенклатура мощностей.

*Таблица 2*

**Сопоставление номенклатур балансов 2009 и 2010 гг.***

| Промышленность | Выведено из номенклатуры по ОКП | Введено в соответствии с ОКВЭД |
|---|---|---|
| Топливная промышленность | 3 | 1 |
| Черная металлургия | 5 | 9 |
| Химическая и нефтехимическая промышленность | 14 | 32 |
| Машиностроение | 68 | 26 |
| Деревообрабатывающая и целлюлозно-бумажная промышленность | 4 | 3 |
| Промышленность строительных материалов и конструкций | 7 | 7 |
| Лёгкая промышленность | 1 | 6 |
| Пищевая промышленность | 4 | 4 |
| Прочие отрасли (Фармацевтическая промышленность) | 9 | 0 |
| Итого | 115 | 88 |

* Данные в табл. 2 рассчитывались по полной БД, включающей экстремальные позиции [7].

Из балансов 2010, 2011, 2012 гг., построенных по системе ОКВЭД, исчезли многие позиции ОКП (115 шт.), и одновременно появились новые (88), которых не было в ОКП (табл. 2).

В процессе очистки информации проверялись: сбалансированность каждой записи, так как нарушение баланса предполагает возможность ошибки; совпадение мощностей при переходе от года к году в динамике; приведение записей с одинаковыми ключами в таблицах разных лет к одним и тем же единицам измерения [6, 7].

После чего все годовые балансовые таблицы были внесены постранично в один файл БД, включающий также ряд примечаний, сводные таблицы сопоставления номенклатур, таблицы расчёта коэффициентов перехода (КП) в разных форматах, и др.

Созданная таким образом на основе исходных балансов Росстата пользовательская пополняемая *исходная БД 1997–2010 гг.*[1] является хорошо структурированной базой, максимально очищенной от возможных ошибок (несогласованности размерностей и номенклатуры, ошибок ввода, сдвига информации, позиций, дающих возможность двойного счёта и др.). На её основе созданы *полная и сокращённая базы данных*. Необходимость создания *сокращённой* БД вызвана тем, что в исходных данных могут появляться отдельные позиции с экстремальной динамикой мощностей, т.е. с резко растущими (в «разы» или даже в «десятки раз»), либо резко снижающимися (вплоть до нуля) их объемами. Эти позиции были условно названы *экстремальными*.

При выбранной методике расчетов обобщающих показателей по промышленности как среднеарифметических невзвешенных величин, эти позиции в *каждом* анализируемом году могут оказывать очень сильное воздействие на уровень средних. После проведения тщательного логического экономического анализа *экстремальные* позиции были удалены из полной базы данных. Полная и сокращённая БД, повторяя структуру исходной базы, содержательно отличаются друг от друга только наполнением.

Несопоставимость БД 1997–2009 гг. (насчитывающей 330 ПМ) с новой БД 2010–2012 гг. (имеющей 359 ПМ)[2] вынудила нас создать ещё одну БД 1997–2012 гг. (насчитывающую 216 ПМ), так называемую сквозную базу данных. То есть в нашем распоряжении есть три динамические БД, и в каждой из них информация согласована, очищена и внутренне не противоречива [6, 7, 8]. БД созданы по одной методике, но структурно отличаются.

Каждый новый очередной БПМ обрабатывается по предложенной методике и записывается в соответствующую БД. Пополнение БД новыми балансами не ограничено.

Процесс создания пользовательских БД во всех случаях, где возможно обойтись без участия человека, автоматизирован.

---

[1] Балансовая таблица 2010 г. в этой БД дана в суженном виде, о чём написано в [7].

[2] Число позиций относится к данным за 2012 г. в БД, из которой не удалены экстремальные позиции [7].

***Анализ погодовой сопоставимости информации.*** Создание структурированных БД существенно облегчило контроль качества данных.

Одним из видов контроля качества данных в динамической БД является анализ информации по сопоставимости ПМ на начало последующего и конец предыдущего года.

Динамическая таблица коэффициентов перехода, построенная по результатам проверки, позволила при экономическом анализе выявить некоторые особенности предлагаемой Росстатом статистики по балансам ПМ[1].

Полная сопоставимость погодовой информации предполагает, что мощность на начало последующего года должна быть равна мощности на конец предыдущего. Отношение этих величин было условно названо коэффициентом перехода – КП и в этих случаях должно быть равно 1 (или 100%).

Там, где этот коэффициент не равен 100%, речь может идти об изменении круга предприятий, относящихся к данной позиции, изменении номенклатуры продукции, просто ошибке, недостоверности информации и др., что, естественно, и приводит к таким показателям. В подобных случаях можно говорить о несопоставимости внутреннего (экономического) содержания записей в БД. В каждом конкретном случае требуется проведение глубокого экономического анализа.

Из приведённой табл. 3 распределения коэффициентов перехода видно, что за период 1997–2009 гг. количество полностью сопоставимых позиций (КП = 1) колеблется в интервале 114–160. Количество частично сопоставимых позиций (КП$\in$ (0,8–1,2)) колеблется от 276 до 319 позиций. Число плохо сопоставимых позиций составляет примерно 4–9% от общего числа позиций, входящих в погодовые БД. Информация 2010 г. по суженной базе данных, как видно из табл. 3, по выбранному критерию достаточно плохо согласуется с остальными данными.

Из табл. 4 видно, что число полностью сопоставимых в динамической БД позиций (КП = 1) колеблется от 36,2 до 48,3% от общего количества рассматриваемых ПМ. Число позиций с КП в полуинтервале $0{,}8 < \text{КП} \leq 1{,}2$ колеблется в интервале 90,8–96,1%.

[1] Дальнейший анализ проводится по информации БД 1997–2009.

*Таблица 3*

**Распределение мощностей по уровню КП в 1998–2010 гг.**

| Интервалы КП | Отношение мощности на начало года к мощности на конец года | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1998/1997 | 1999/1998 | 2000/1999 | 2001/2000 | 2002/2001 | 2003/2002 | 2004/2003 | 2005/2004 | 2006/2005 | 2007/2006 | 2008/2007 | 2009/2008 | 2010/2009 |
| | Количество производственных мощностей | | | | | | | | | | | | |
| $0 < x \leq 0{,}8$ | 4 | 6 | 6 | 8 | 4 | 5 | 6 | 9 | 3 | 9 | 11 | 7 | 23 |
| $0{,}8 < x < 1$ | 65 | 52 | 75 | 77 | 88 | 83 | 88 | 89 | 54 | 71 | 89 | 70 | 54 |
| $x = 1$ | 110 | 142 | 114 | 115 | 139 | 153 | 153 | 148 | 145 | 149 | 135 | 159 | 35 |
| $1{,}0 < x \leq 1{,}2$ | 101 | 95 | 97 | 99 | 77 | 84 | 75 | 73 | 109 | 90 | 82 | 86 | 80 |
| $1{,}2 < x$ | 24 | 8 | 8 | 7 | 16 | 8 | 11 | 13 | 21 | 11 | 10 | 7 | 34 |
| Итого | 304 | 303 | 300 | 306 | 324 | 333 | 333 | 332 | 332 | 330 | 327 | 329 | 226 |

*Таблица 4*

**Распределение мощностей по уровню КП в 1998-2010 гг., %**

| Интервалы КП | Отношение мощности на начало года к мощности на конец года, % | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1998/1997 | 1999/1998 | 2000/1999 | 2001/2000 | 2002/2001 | 2003/2002 | 2004/2003 | 2005/2004 | 2006/2005 | 2007/2006 | 2008/2007 | 2009/2008 | 2010/2009 |
| | Количество производственных мощностей, % | | | | | | | | | | | | |
| $0 < x \leq 0{,}8$ | 1,3 | 2,0 | 2,0 | 2,6 | 1,2 | 1,5 | 1,8 | 2,7 | 0,9 | 2,7 | 3,4 | 2,1 | 10,2 |
| $0{,}8 < x < 1$ | 21,4 | 17,2 | 25,0 | 25,2 | 27,2 | 24,9 | 26,4 | 26,8 | 16,3 | 21,5 | 27,2 | 21,3 | 23,9 |
| $x = 1$ | 36,2 | 46,9 | 38,0 | 37,6 | 42,9 | 45,9 | 45,9 | 44,6 | 43,7 | 45,2 | 41,3 | 48,3 | 15,5 |
| $1{,}0 < x \leq 1{,}2$ | 33,2 | 31,4 | 32,3 | 32,4 | 23,8 | 25,2 | 22,5 | 22,0 | 32,8 | 27,3 | 25,1 | 26,1 | 35,4 |
| $1{,}2 < x$ | 7,9 | 2,6 | 2,7 | 2,3 | 4,9 | 2,4 | 3,3 | 3,9 | 6,3 | 3,3 | 3,1 | 2,1 | 15,0 |
| Всего | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | **Отношение позиций с КП, входящих в интервалы ($0 \leq x \leq 0{,}8$): ($1{,}2 < x$) к общему числу позиций, %** | | | | | | | | | | | | |
| Отношение | 9,2 | 4,6 | 4,7 | 4,9 | 6,2 | 3,9 | 5,1 | 6,6 | 7,2 | 6,1 | 6,4 | 4,3 | 25,2 |

Это существенно больше, но при этом надо понимать, что и здесь внесена некоторая систематическая ошибка. А величина отношения числа ПМ, у которых КП входит в указанный выше полуинтервал, к общему числу ПМ и колеблется от 3,9 до 9,2%, говорит, по-видимому, о том, что такие позиции в расчёты средних показателей по промышленности включать не стоит, так как они могут значительно исказить результат анализа. Эти рассуждения относятся к периоду 1997–2009 гг. При сопоставлении 2009 и 2010 гг. КП оказалось, что более 25% позиций несопоставимы или сопоставимы условно. Информация 2010 г. относится к суженной БД (таб. 5).

По величине среднеквадратического отклонения $\sigma(x)$ можно судить о степени разброса значений в рассматриваемом множестве. Чем меньше эта величина, тем более плотно значения во множестве сгруппированы около средней величины. В общем смысле среднеквадратическое отклонение можно считать мерой неопределенности. Оно очень важно для определения правдоподобности изучаемого явления в сравнении с предсказанным теорией (или предполагаемой гипотезой) значением: при большом значении среднеквадратического отклонения полученные значения или метод их получения следует перепроверить. Например, в физике среднеквадратическое отклонение используется для определения погрешности серии последовательных измерений какой-либо величины.

По коэффициенту вариации $v(x)$ случайной величины (мера относительного разброса случайной величины) можно судить о том, какую долю среднего значения этой величины составляет её средний разброс.

Анализ значений $\sigma(x)$ и $v(x)$ в табл. 5 показывает, что при проведении логического контроля надо обратить пристальное внимание на информацию. Особенно это относится к сопоставимости по КП 2000 и 2001 гг., 2003 и 2004 гг., 2009 и 2010 гг.

Хотя говорить о полной внутренней (экономической) сопоставимости ряда позиций нельзя, мы оставили эти и подобные им записи в БД (КП, не равных 100% при рассмотрении динамических рядов отношений ПМ по всем позициям оказалось много), так как была предложена методика оценки динамических характеристик, которая устраняет этот эффект.

Но неравенство коэффициентов перехода 100% может говорить и о недостоверности информации.

**Статистические характеристики сопоставимости коэффициентов перехода** *Таблица 5*

| Отношение мощности на начало года к мощности на конец года, % | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Пока-затель | 1998/1997 | 1999/1998 | 2000/1999 | 2001/2000 | 2002/2001 | 2003/2002 | 2004/2003 | 2005/2004 | 2006/2005 | 2007/2006 | 2008/2007 | 2009/2008 | 2010/2009 |
| M(*x*) | 105,9 | 101,7 | 101,0 | 105,8 | 105,0 | 100,5 | 120,4 | 102,4 | 104,3 | 103,7 | 102,4 | 101,5 | 167,9 |
| σ(*x*) | 29,9 | 16,8 | 15,9 | 73,6 | 32,2 | 16,0 | 274,6 | 23,0 | 18,9 | 32,1 | 35,3 | 22,0 | 550,2 |
| *v*(*x*) | 28,2 | 16,5 | 15,7 | 69,5 | 30,6 | 16,0 | 228,0 | 22,4 | 18,1 | 30,9 | 34,5 | 21,7 | 327,6 |

**Некоторые статистические характеристики полной и сокращённой БД** *Таблица 6*

| Показатель | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Число позиций ПМ** | | | | | | | | | | | | | |
| БД (полная) | 305 | 304 | 304 | 307 | 324 | 334 | 335 | 335 | 331 | 332 | 332 | 329 | 330 | 229 |
| БД (сокращённая) | 301 | 299 | 299 | 303 | 318 | 329 | 332 | 330 | 326 | 331 | 327 | 326 | 326 | 229 |
| | **Статистические параметры для полной БД, %** | | | | | | | | | | | | | |
| M[*x*] | 97,1 | 94,5 | 98,1 | 100,4 | 98,5 | 99,5 | 99,2 | 96,5 | 103 | 100 | 104,8 | 103,7 | 96,7 | 100,4 |
| σ[*x*] | 13,2 | 13,9 | 24,1 | 32,5 | 16,2 | 20,4 | 18,6 | 20,7 | 53,8 | 27,3 | 74,9 | 52,7 | 42,9 | 15,8 |
| *v*[*x*] | 13,5 | 14,7 | 24,5 | 32,4 | 16,4 | 20,5 | 18,8 | 21,5 | 52,2 | 27,3 | 71,4 | 50,8 | 44,4 | 15,7 |
| | **Статистические параметры для сокращённой БД, %** | | | | | | | | | | | | | |
| M[*x*] | 97,4 | 95,3 | 96,8 | 99 | 98,4 | 99,5 | 99,4 | 97,2 | 99,6 | 100,3 | 101,1 | 101,5 | 95,5 | 100,4 |
| σ[*x*] | 12,7 | 11,4 | 14,2 | 15,7 | 16,2 | 19,2 | 18,6 | 19,3 | 22,5 | 26,8 | 20,2 | 26,3 | 20,6 | 15,8 |
| *v*[*x*] | 13,1 | 12 | 14,7 | 15,9 | 16,5 | 19,3 | 18,7 | 19,9 | 22,6 | 26,7 | 19,9 | 25,9 | 21,5 | 15,7 |
| | **Разности параметров** | | | | | | | | | | | | | |
| Число удалённых экстремальных позиций | 4 | 5 | 5 | 4 | 6 | 5 | 3 | 5 | 5 | 1 | 5 | 3 | 4 | 0 |
| $M[x]_п$-$M[x]_с$ | –0,3 | –0,8 | 1,3 | 1,4 | 0,1 | 0 | –0,2 | –0,7 | 3,4 | –0,3 | 3,7 | 2,2 | 1,2 | 0 |
| $σ[x]_п$-$σ[x]_с$ | 0,5 | 2,5 | 9,9 | 16,8 | 0 | 1,2 | 0 | 1,4 | 31,3 | 0,5 | 54,7 | 26,4 | 22,3 | 0 |
| $v[x]_п$-$v[x]_с$ | 0,4 | 2,7 | 9,8 | 16,5 | -0,1 | 1,2 | 0,1 | 1,6 | 29,6 | 0,6 | 51,5 | 24,9 | 22,9 | 0 |

***Фактор неопределённости в исходной информации***. В связи с тем, что в исходной информации могут быть ошибки, а некоторые позиции после логического контроля признаны нами как экстремальные и удалены (тем самым внесён элемент *волюнтаризма*, хоть и обоснованного), результаты расчётов средних показателей искажаются.

Предположим, что в каждом исходном числе присутствует фактор неопределённости. Попытаемся оценить достоверность и устойчивость полученных результатов расчёта на примере темпов роста в динамике по полной и сокращённой БД как для вероятностного, так и для статического вариантов [9, 10].

Полученные предварительные данные, с одной стороны, подтверждают результаты проведенного выше экономического анализа, с другой – говорят о том, что выходные параметры по средним показателям соответствуют истине с определённой долей вероятности. Чтобы получить количественную оценку достоверности, надо провести ещё ряд исследований.

В табл. 6 приведены некоторые статистические оценки результатов расчётов по полной и сокращённой БД для статического варианта.

По изменению среднеквадратичных ошибок в каждом году динамического ряда (см. табл. 6) можно судить, что при удалении экстремальных позиций уменьшается разброс значений в выборке, т.е. они более плотно группируются около среднего[1].

Но если разности математических ожиданий выборок при удалении позиций изменяются по абсолютной величине в пределах 0–3,7% (M[*x*] совпадает со среднеарифметической невзвешенной величиной в принятой ранее постановке), то разности среднеквадратичных ошибок меняются в пределах от 0–54,7%. Причём это никак не связано с количеством удаляемых позиций. То же относится к коэффициентам вариации.

Так, если сравнить, например, 2001 г. и 2007 г., где было удалено примерно одинаковое количество позиций, то видна огромная разница в разбросе значений выборки относительно среднего.

Если же предположить, что каждое из исходных данных является случайным числом, распределённым по нормальному закону, то можно провести по правилу $3\sigma$[2] анализ информации. Он показывает, что

[1] Более подробный анализ статистических оценок полученных результатов пока не проводился.

[2] "3σ" – правило, по которому значение нормально распределённой случайной величины лежит в интервале $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$ не менее чем с 99,7%-й достоверностью (при условии, что величина $\bar{x}$ истинная, а не полученная в результате обработки

достоверной (расчёт проводился по сокращённой БД) является от 96,2–99,4% информации. Показатели расчёта достоверности информации (%) по "правилу 3σ" выглядят следующим образом:

| 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 98,7 | 98,3 | 97,0 | 98,7 | 96,2 | 98,2 | 97,9 | 97,6 | 99,1 | 97,3 | 97,6 | 99,4 | 97,2 | 97,4 |

В связи с тем что (см. выше) агрегированные данные по каждому атрибуту производственных мощностей, которые нам даёт Росстат, несут в себе некоторую вероятность ошибки, можно предположить, что каждое значение ПМ является случайной величиной, распределённой в некоторой области существования $a_{ij} \pm \alpha * a_{ij}$ по (для простоты расчёта) равномерному закону, хотя можно выбирать и другой. Подробно обоснованность постановки и решения задач с фактором неопределённости в исходных параметрах изложена в [9, 10, 11]. Здесь же мы выясним, как влияет фактор неопределённости в исходной информации балансов ПМ на поведение статистических параметров выборки.

Полагая, что значение ПМ лежит в интервале, равном (ПМ – L*ПМ, ПМ + L*ПМ), где L равно длине полуинтервала, выраженной в процентах, вычисляем левую границу интервала для каждого значения ПМ, и, пользуясь методом Монте-Карло, при помощи датчика случайных чисел, равномерно распределённых на интервале (0,1), получаем матрицу случайных ПМ. Одновременно вычисляем случайное значение темпов роста по каждой позиции и по промышленности в целом по каждому году, их математические ожидания, среднеквадратические ошибки и коэффициенты вариации.

При проведении серии испытаний значения L принимаем равными 5, 15, 25%, и смотрим, насколько устойчиво ведут себя статистические параметры выборки по каждому году. В качестве датчика взят *генератор* равномерно распределённых случайных чисел. Он выбран потому, что *программные* датчики случайных чисел дают большую повторяемость значений.

На рис. 1, 2, 3 приведены графики статистических параметров, полученных в результате эксперимента на основе расчётов темпов роста мощностей (ТМП), полученных по полной БД.

---

выборки). В нашем случае, поскольку $\bar{x}$ истинная неизвестна, правило превращается в правило (3s) [12].

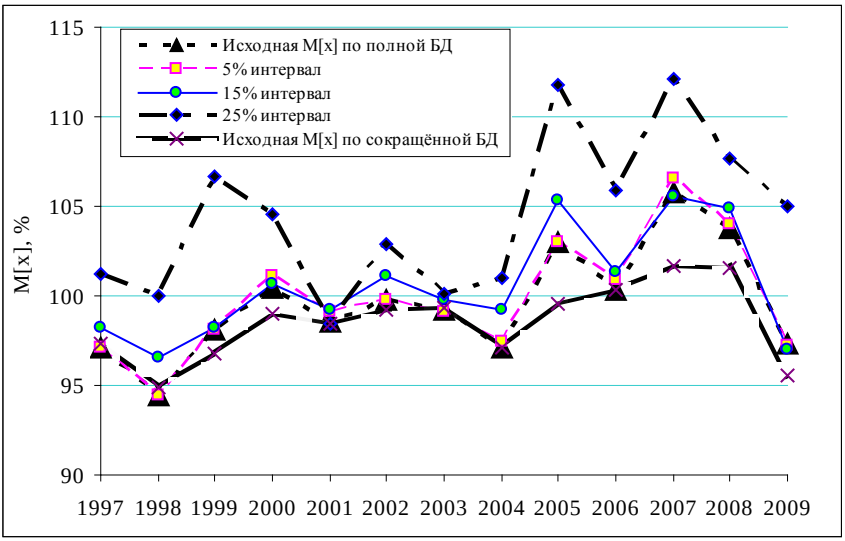


*Рис. 1.* Влияние интервала разброса исходной информации на погодовые значения M[x]

Изображённые на рисунке 1 статические (исходные) погодовые значения M[x], рассчитанные по полной и сокращённой базам данных, и M[x], полученные в результате расчётов с применением метода статистических испытаний Монте-Карло, показывают, что при внесении элемента случайности в исходные данные, тенденции поведения параметров выходных величин сохраняются.

Величина выходных параметров довольно слабо зависит от интервала разброса исходных данных (X) в пределах интервалов X ± 10%. То есть вероятностный характер информации, независимо от величины области распределения случайной величины, говорит об устойчивости системы.

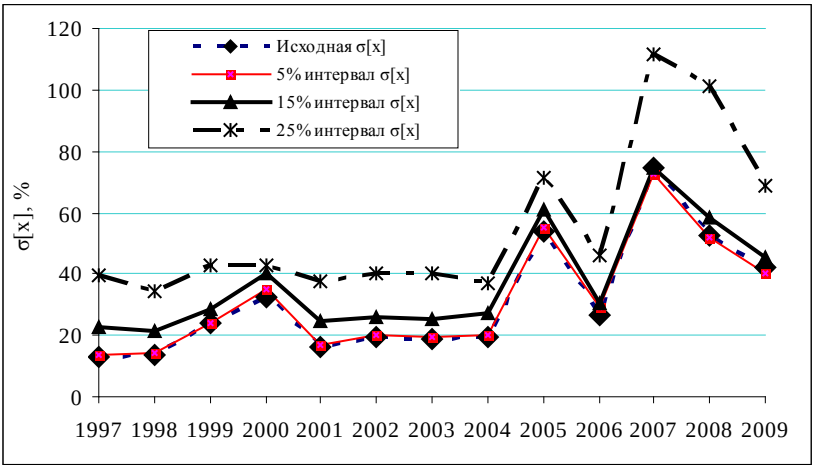


*Рис. 2.* Влияние интервала разброса исходной информации на погодовые значения σ[*x*], %

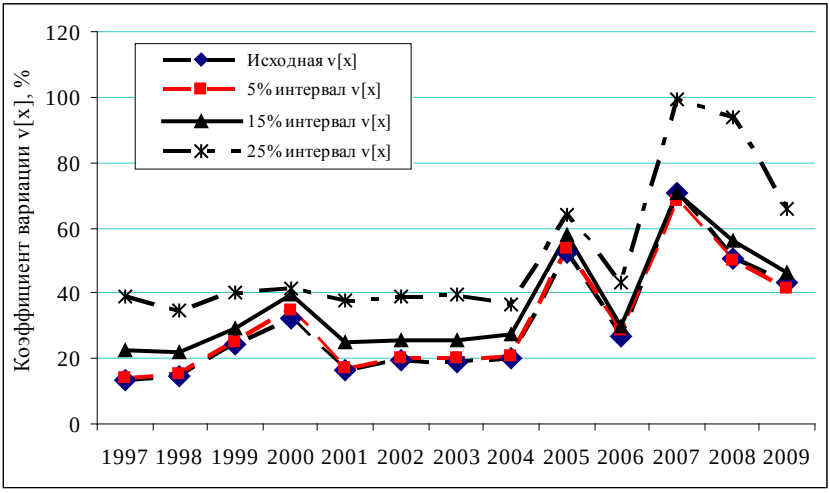


*Рис. 3.* Коэффициент вариации *v*[*x*] случайных величин

Присутствующие в информации системные ошибки, привнесённые в неё по тем или иным причинам, а также агрегация, удаление

экстремальных позиций не изменяют тенденцию, хотя в каждом конкретном статическом случае на неё влияют. Аналогичный вывод можно сделать относительно тенденций среднеквадратичной ошибки и коэффициентов вариации.

*Из сказанного выше можно сделать только предварительные выводы*, так как работа требует ряда дополнительных статистических оценок и проверки гипотез.

Хотя при подготовке информации к расчётам проводились все виды контроля[1], ни один из них не даёт количественный ответ на вопрос, насколько те или иные действия при очистке информации вносят искажения в результаты обработки данных при дальнейшем анализе.

Анализ сопоставимости объёмов мощностей на конец предыдущего и начало последующего года показал, что 3–9% позиций несопоставимы по этому признаку и нуждаются в тщательном экономическом и логическом анализе.

Выяснилось также, что даже после отбраковки экстремальных позиций, достоверными остаются только 96,2–99,4% информации.

Проведённый с применением метода статистических испытаний Монте-Карло эксперимент показал, что неопределённость в исходных данных не меняет тенденции среднего по промышленности индекса мощностей, хотя в каждом конкретном статическом случае существенно влияет на его уровень.

[1] Этот процесс достаточно подробно освещён в отчётах и публикациях, посвящённых данной тематике [6, 7].

## Литература


1. Щиголев Б.М. Математическая обработка наблюдений. – М.: Государственное издательство физико-математической литературы, 1962. – 344 с.
2. Неганова Л.М. Статистика, конспект лекций. – М.: Институт экономики и права, Изд-во Юрайт, 2010. – 220 с.
3. Неганова Л.М. Общая теория статистики / Учеб. пособие. – М.: Изд-во РИОР, 2007. – 96 с.
4. Автоматизированные информационные технологии в экономике / под ред. проф. Г.А. Титоренко. – М.: Компьютер, ЮНИТИ, 1998. – 400 с.
5. Компьютерные технологии обработки информации / под ред. С.В. Назарова. – М.: Финансы и статистика. – 1995. – 30 с.
6. Сатанова Э.А. Банк данных промышленных мощностей Российской Федерации (1997–2008 гг.) // Отраслевой и макроэкономический аспекты развития российской экономики. – Новосибирск, 2010. – С. 209–224.
7. Селивёрстова Н.Н., Сатанова Э.А. Анализ последствий экономического кризиса 2008–2009 гг. для развития и использования мощностей промышленности РФ / ИЭОПП СО РАН. Рукопись отчёта по плану НИР. – Новосибирск, 2012. – 82 с.
8. Карпов В.Э., Карпова И.П. Об одной задаче очистки и синхронизации данных // Информационные технологии. – 2002. – № 9.
9. Фактор неопределённости в межотраслевых моделях // Березин С.А., Лавровский Б.Л., Рыбакова Т.А., Сатанова Э.А. – Новосибирск: Наука, 1983. – 125 с.
10. Сатанова Э.А. Экспериментальное исследование вероятностной полудинамической модели планового межотраслевого баланса производственных мощностей // Проблемы моделировании народного хозяйства. Ч. III. – Новосибирск, 1973.
11. Сатанова Э.А. Исследование полудинамической модели баланса производственных мощностей в вероятностной постановке методом статистической имитации // Оптимизационные и балансовые модели народного хозяйства. – Новосибирск, 1977. – С. 142–160.
12. Теория вероятностей и математическая статистика: Учеб. пособие для вузов / Гмурман В.Е. – 9-е изд. – М.: Высшая школа, 2003. – 479 с.