УДК 621.3:311.213:001.891.573

ОСОБЕННОСТИ СБОРА И ОБРАБОТКИ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МОДЕЛЕЙ ЭНЕРГОПОТРЕБЛЕНИЯ

А.А. Капанский, Е.Л. Шенец

Гомельский государственный технический университет имени П. О. Сухого (Беларусь)

ОАО «Газпром трансгаз Беларусь»

Аннотация. Статья посвящена проблеме сбора и обработки статистических данных, которые являются основой построения вероятностно-статистических моделей энергопотребления. Приведен перечень минимального объема данных, который включает в себя статистику годовых затрат энергоресурсов по статьям расхода; объемы производственной деятельности; перечень выполненных и планируемых энергосберегающих мероприятий.

В статье приводятся методы восстановления данных при наличии пропусков и методы снижения размера первичной информации. Для уменьшения объема данных рассмотрен метод отбора показателей наиболее весомой группы факторов, воздействующих на энергоэффективность производства, и метод снижения размерности данных в разрезе отраслевого прогнозирования, за счет выбора наиболее энергоемких промышленных предприятий.

Ключевые слова: математическое моделирование, база данных, энергоэффективность, энергоресурсы, мероприятия по энергосбережению.

Введение

Математическое моделирование режимов потребления (топливно-энергетических ресурсов) ТЭР является одним из важнейших элементов оценки и прогнозирования энергоэффективности производства [1, 2, 3]. Исследование функциональных связей между энергопотреблением и воздействующими факторами основано на обработке информационной

базы данных (ИБД) методами статистического анализа. К основным этапам построения моделей следует отнести: формирование ИБД об объекте исследования; выбор оптимального метода построения модели; обработка данных; выполнение требуемых расчетов; оценка погрешности (верификация) модели и анализ полученных результатов.

В рассматриваемой статье авторы ставят перед собой цель исследовать особенности сбора и обработки статистических данных, которые являются основой построения многофакторных регрессионных моделей, описывающих режимы потребления ТЭР.

Снижение размерности массива статистических данных

При выборе вероятностно-статистической модели в поле зрения инженера попадают две совокупности объектов: реальная сформированная информационная выборка и теоретически-представляемый набор называемых данных генеральной совокупности. Основные свойства исследуемой выборки могут быть определены по имеющимся статистическим данным, в то время как свойства генеральной совокупности устанавливаются исходя эмпирических свойств выборки. В результате исследования особенностей выборочных данных появляется возможность определить необходимый вид модели и методы обработки ИБД. Выделяют следующие типы выборочных данных [4]:

1. Пространственная выборка. Данные приведенной выборки не имеют временного фактора и могут быть представлены в виде:

$$X_p = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{vmatrix}, \quad (1)$$

где X_p – массив информационной базы данных; p – количество подобъектов исследования (промышленных предприятий в рамках отрасли); x_{ij} – результат i-го наблюдения j-го объясняющего фактора; n – количество переменных (факторов); m – количество наблюдений.

2. Временной ряд. В статистике временных наблюдений важен порядок следования анализируемых данных. Математическая модель, сформированная с использованием временных рядов, представляется зависимостью результирующего признака от воздействующих переменных, одним из которых выступает фактор времени t.

Приведенные типы выборочных данных определяют вид итоговой математической модели:

1. Регрессионные модели. Такой вид моделей описывается функциональной связью между объясняющими факторами и зависимой переменной:

$$y = f(x_1, x_2, \ldots, x_n, b_1, b_2, \ldots, b_n) + \varepsilon, \quad (2)$$

где b_n – коэффициент регрессии *j*-го фактора; ε – случайная составляющая модели.

2. Модели временных рядов. Здесь выделяются модели тренда и сезонности. Трендовая модель описывается уравнением:

$$y(t) = f(t) + \varepsilon = b_0 + b_1 t + \varepsilon, \quad (3)$$

где b_0 – свободный член модели; b_1 – коэффициент регрессии перед фактором времени.

3. *Модель сезонности* включает в себя периодическую функцию $S(t)$:

$$y(t) = S(t) + \varepsilon. \quad (4)$$

Конечный вид модели определяется целями исследования, составом и объемом исходной информации. Для поставленных в работе задач, связанных с оценкой и прогнозированием энергоэффективности производства, могут использоваться как регрессионные, так и временные модели, кроме того и их сочетание.

Детальный анализ входной информации исследуемых промышленных предприятий приводит к необходимости накопления большого объема выборочных данных. Набор первичной информации для построения моделей энергопотребления *i*-го промышленного предприятия определяется спецификой производства, заданной точностью решения поставленной задачи и требуемым объемом выборки. Выделяя общие признаки исследуемых предприятий, минимальный набор данных должен определяться:

– *статистикой годовых затрат энергоресурсов по статьям расхода (топливо, тепло- и электроэнергия);*

– *статистикой объемов производственной деятельности, сопоставимой с принятым уровнем дискретизации;*

– *перечнем выполненных и планируемых энергосберегающих мероприятий.*

Не включенные в первичный набор данных факторы, определяются различными техническими соображениями, одним из которых может выступать громоздкость исходной информации. Одной из центральных проблем статистического анализа является проблема снижения размерности ИБД [5]. В соответствии с этой проблемой предполагается, что существует возможность лаконичного описания исследуемой зависимой переменной за счет использования в качестве входной матрицы наиболее информативных факторов, что способствует сокращению объема выборки. Для сокращения объема первичной информации могут использоваться следующие методы [6]:

1. Отбор показателей наиболее значимой группы факторов в соответствии с условиями решаемых задач. Рассмотренный метод, определяет снижение размерности выборки за счет уменьшения количества предикторов $x_1, x_2 \dots x_n$, учитывая специфику решаемой задачи. К примеру, для удовлетворения прогноза энергопотребления промышленного объекта во многих случаях нет необходимости учитывать абсолютно всю группу факторов, которые описывают модель с коэффициентом детерминации близким к единице ($R^2 \to 1$). Увеличение объема предикторов в ряде случаев способствует только усложнению процедуры сбора и обработки данных. При этом точность прогноза может не существенно отличаться от требуемой. Таким образом, в соответствии с рассмотренным подходом реальное количество регрессоров заменяется меньшим с условием удовлетворения критерия информативности (в форме записи *i*-е наблюдения опущены):

$$X_p = (x_1, x_2 \dots . x_n), \tag{5}$$

$$X'_p = \left(x_1, x_2 \dots . x_l \,\middle|\, l < n, I \geq I_{\text{зад}}\right), \tag{6}$$

где X_p, X'_p – множество предикторов до и после снижение размерности соответственно; p – количество исследуемых множеств (подобъектов системы); n, l – количество предикторов до и после снижения размерности; I, $I_{\text{зад}}$ – фактический и заданный критерий информативности, например точность прогноза.

2. Сжатие массивов статистических данных за счет выбора наиболее значимых подобъектов. При большом объеме не только факторных, но и критериальных (зависимых) переменных, к примеру, когда обработка данных и задача прогноза энергопотребления затрагивает не одно конкретное предприятие, а отрасль промышленности, обработка *p*-мерного массива данных многократно усложняется. Для снижения массива информационной базы данных без значительной потери качества итоговой отраслевой модели наряду с регрессионным анализом используется кластерный, который позволяет классифицировать объект по степени значимости входящих в него множеств. Такой поход позволяет снизить размерность исходной выборки включающей в себя p подобъектов, до k-мерной эталонной (наиболее значимой) выборки ($k << p$). Тогда при сочетании метода *отбора показателей наиболее значимой группы факторов в соответствии с условиями решаемых задач* и метода *сжатия массивов статистических данных* итоговое множество данных снижается с X до X':

$$X = \{X_1, X_2 \ldots, X_p\}, \quad (7)$$

$$X' = \{X_1', X_2' \ldots_1, X_k'\}. \quad (8)$$

3. Наглядное представление статистических данных. Сущность метода заключается в визуальном анализе расположения данных на пространственной плоскости в результате чего, появляется возможность увидеть сгустки информации и соответственно выявить наиболее значимую группу наблюдений или наиболее значимые факторы с точки зрения прогнозирования. Использования рассмотренного метода ограничивается размерностью видимого пространства (анализ не более чем в трех пространственных плоскостях).

Структура ИБД в рамках отдельного предприятия и отрасли промышленности может быть представлена на рисунке 1.

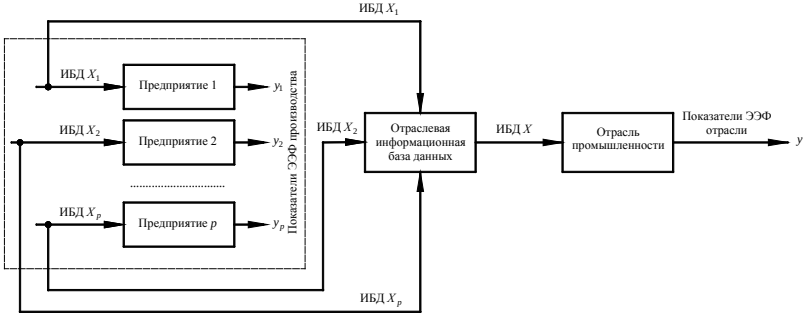

Рисунок 1 – Структура информационной базы статистических данных объекта исследования

Обработка массива данных содержащего пропуски

Обратной стороной увеличения размерности информационной базы данных является её низкая наполняемость и наличие пропусков [7]. В практических условиях основной причиной формирования пустот в информационной матрице является некорректная работа автоматизированных систем управления технологическим процессом (АСУТП). Условно АСУТП следует разделить на три уровня: нижний, средний и верхний. *Нижний уровень* представляет собой совокупность цифровых и аналоговых датчиков, предназначенных для непосредственного измерения и контроля состояния технологического оборудования. *Средний уровень* – программируемые логические контроллеры с необходимым набором модулей, реализующие систему сбора данных от устройств нижнего уровня и выработку управляющих сигналов на основе данных полученных из верхнего уровня. *Верхний уровень* –

автоматизированные рабочие места (АРМ) диспетчеров и сетевое коммутационное оборудование, обеспечивающие сбор данных с устройств среднего уровня, сбор данных для создания архива параметров и состояний работы оборудования, архива аварийный ситуаций, действий диспетчера. Уязвимость процессов сбора, обработки и передачи данных на каждом рассмотренном уровне приводит к высокой степени вероятности появления промахов или пустот даже при небольшом составе факторов.

В практике обработки данных с пропусками следует выделить следующие методы:

1. Удаление строки матрицы ИБД содержащей пустоты. Такой метод является наиболее простым в использовании и позволяет сформировать базу данных содержащую полный набор исходной информации. Однако такой подход может привести к ситуации, в которой минимальный объем выборки не удовлетворяет условиям построения качественной модели. К примеру, в [4] указывается, что для линейной модели регрессии число наблюдений m должно быть не менее чем в 7 раз больше числа факторов модели x_n. Таким образом, исключение пустот в ИБД путем удаления строк, не содержащих информацию, может существенно исказить итоговую модель и привести к некорректной оценки и интерпретации полученных результатов. Использовать рассмотренный метод имеет смысл только в том случае, если удаление пустот в информационной матрице данных сопровождается снижением её размерности.

2. Заполнение пропущенных данных. Использование данного метода изначально имеет преимущества перед рассмотренным ранее, так как предотвращает потерю первичной информации. Выбор подхода, по которому заполняются пустоты, зависит от различных факторов, например, таких как причины формирования пропусков, вариация исходных данных и др.[8]. Выделяют следующие методы [9]:

а) Метод среднего значения известной выборки. Самый простой метод заполнения пропущенных данных $x_{\text{проп}i}$, который заключается в замене пустот средним арифметическим значением выборки:

$$x_{\text{проп}i} = \sum x_i / m, \qquad (9)$$

где x_i – результат i-го наблюдения; m – объем известных данных выборки.

Недостатком является усреднение данных с большим разбросом или, к примеру, данных характеризующих сезонность наблюдений. Для наглядности рассмотрим пример, в котором одним из исследуемых факторов являлась температура. В результате наблюдений был потерян набор данных за два характерных месяца

года – июнь и июль. На рисунке 2 визуально наблюдаются значительные расхождения реальных данных со средним. Относительная погрешность замены составила 71 %.

Рисунок 2 – Замена пропущенных данных методом подстановки среднего значения выборки

б) Метод замены пропусков средним значением N соседних точек. Метод предполагает определение среднего значения по обе стороны от пропуска. Отличие от предыдущего метода заключается в том, что для определения среднего арифметического не используются все значения ИБД, а *N* ближайших значений:

$$x_{\text{проп}i} = \sum x_i / N, \qquad (10)$$

где x_i – результат *i*-го наблюдения в области принятого объема данных усреднения; *N* – принятый объем данных усреднения.

Рисунок 3 – Замена пропущенных данных методом подстановки среднего значения соседних точек

К примеру, если $N = 2$ – используется всего два значения слева и справа от пропуска. Результат использования рассмотренного метода приведен на рисунке 3. Погрешность определения данных составила – 12%.

в) Метод линейной интерполяции соседних данных статистики. Данный метод по своей сути представляет собой замену пропусков прямой линией между известными величинами. На рисунке 4 приведен результат замены пропусков линейной интерполяцией.

Поиск неизвестных данных осуществляется по формуле:

$$x_{\text{проп}i} = x_{i-1} + \frac{x_{i+1} - x_{i-1}}{n_{i+1} - n_{i-1}}\left(m_i - m_{i-1}\right), \qquad (11)$$

где x_{i-1} – значение наблюдения предшествующее $x_{\text{проп}i}$; x_{i+1} – следующее за $x_{\text{проп}i}$ известное наблюдение; m_i – номер пропуска в матрице данных; m_{i-1}, m_{i+1} – предыдущее и следующее известное значение номера.

Рисунок 4 – Замена пропущенных данных методом линейной интерполяции

В случае, когда неизвестна тенденция пропущенной выборки, метод может оказаться достаточно точным в использовании. На рассматриваемом примере погрешность заполнения данных существенно снизилась по сравнению с предыдущими методами и составила -4 %.

г) Метод замены пропусков линейным трендом. Метод эффективен при наличии автокорреляции в рядах динамики, то есть когда наблюдается связь между последовательностями величин одного

ряда в виде строго убывающей или возрастающей тенденции. В таком случае отсутствующие данные заменяются предсказанной линией регрессии:

$$x_{\text{проп}i} = b_0 + b_1 m_i. \quad (12)$$

При замене пропусков линией тренда в матрице данных имеющий сезонную составляющую без заметной динамики роста или спада результат существенно расходится с реальными данными (см. рисунок 5). Погрешность рассмотренного примера составила 71 %.

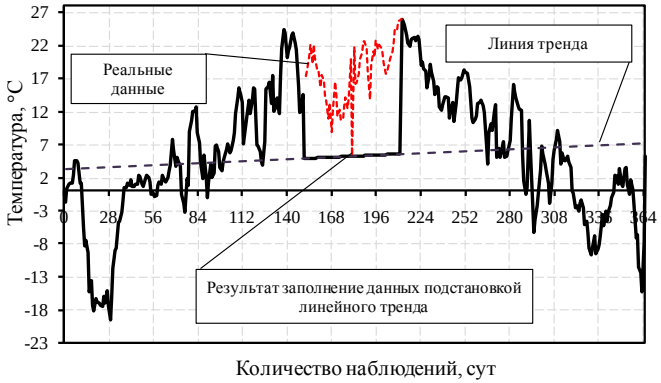

Рисунок 5 – Замена пропущенных данных методом подстановки линейного тренда

Выводы

1. Минимальный набор данных для построения моделей энергопотребления и прогнозирования энергоэффективности производства должен содержать: статистику годовых затрат энергоресурсов по статьям расхода; статистику объемов производственной деятельности; перечень выполненных и планируемых энергосберегающих мероприятий.

2. Для снижения громоздкости информационной базы данных используют метод отбора показателей наиболее весомой группы факторов, воздействующей на энергоэффективность производства, и метод снижения размерности данных в объеме отраслевого прогнозирования, за счет выбора наиболее энергоемких промышленных предприятий.

3. Выбор метода восстановления данных должен основываться на следующих соображениях:

– *метод среднего по известной выборке и N значений соседних точек* следует использовать при небольшой дисперсии наблюдений, отсутствии сезонности и ярко выраженной динамики;

– *метод линейной интерполяции соседних данных статистики* наиболее эффективен при наличии сезонного фактора в данных;

– *метод замены пропусков линейным трендом* наиболее эффективен при наличии автокорреляции в рядах динамики.

Список использованных источников:

1. Показатели энергоэффективности: основы статистики: International Energy Agency. – France. 2014.

2. Токочакова, Н.В. Расчетно-статистические модели режимов потребления электроэнергии как основа нормирования и оценки энергетической эффективности / Н.В. Токочакова, Д.Р. Мороз // Минск: «Энергоэффективность», №1, 2006. – с. 14–15., №2, 2006. – С. 14-15.

3. Токочакова Н.В. Управление энергоэффективностью промышленных потребителей на основе моделирования режимов электропотребления / Н.В. Токочакова // Известия высших учебных заведений и энергетических объединений СНГ. Энергетика. 2006. № 3. С. 67–75.

4. Воскобойников, Ю. Е. Эконометрика в Excel : учеб. пособие. В 2 ч. Ч. 1. Парный и множественный регрессионный анализ / Ю. Е. Воскобойников ; Новосиб. гос. архитектур.-строит. ун-т. – Новосибирск: НГАСУ (Сибстрин), 2005. – 182 с.

5. Айвазян, С. А. Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. – М.: ЮНИТИ, 1988. – 1005 с.

6. Айвазян, С. А. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; под ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.: ил.

7. Радчикова, Е. С. Анализ применения способов заполнения пропусков в данных во временных рядах в экологических исследованиях / Е. С. Радчикова // Экология и защита окружающей среды : сб. тез. докл. Междунар. науч.-практ. конф., 19–20 марта 2014 г. – Минск, 2014. – С. 112 – 116.

8. Мальцев, К.А. / Статистический анализ данных в экологии и природопользовании / К. А Мальцев, С.С. Мухарамова // Казань: Казанский (Приволжский) федеральный университет, 2011 г. – 50 с.

9. Злоба, Е. / Статистические методы восстановления пропущенных данных/ Е. Злоба, И. Яцкив // Рига: Институт транспорта и связи, 2002 г. – 45 с.

Капанский Алексей Александрович, магистр, Беларусь, Гомельский государственный технический университет им. П.О.

Сухого,
kapanski@mail.ru.

Шенец Евгений Леонидович, магистр, Беларусь, начальник отдела капитального строительства объектов энергетики ОАО «Газпром трансгаз Беларусь»

FEATURES OF DATA COLLECTION AND PROCESSING FOR CONSTRUCTION OF PROBABLE-STATISTICAL MODELS OF ENERGY CONSUMPTION

A.A Kapanskiy, E.L. Shenets
Gomel State Technical University named after P.O. Sukhoi (Belarus)
Gazprom transgaz Belarus

Abstract. The article is devoted to the collection and processing of statistical data, which are the basis for constructing probability-statistical models of energy consumption. A list of the minimum amount of data is given, which includes statistics of annual energy costs by expense items; Volumes of production activity; List of implemented and planned energy-saving measures. The article presents the methods of data recovery in the presence of gaps and methods to reduce the size of primary information. To reduce the amount of data, a method for selecting the indicators of the most significant group of factors affecting the energy efficiency of production, and a method for reducing the size of data in the context of sectoral forecasting, by selecting the most energy-intensive industrial enterprises is viewed.
Keywords: Mathematical modeling, database, energy efficiency, energy, energy saving measures.