Алгоритмическое и программное обеспечение

УДК 517.4:519.652

К ВОПРОСУ ВОССТАНОВЛЕНИЯ УЧЕТНЫХ ДАННЫХ НА ХИМИЧЕСКИХ ПРЕДПРИЯТИЯХ

Волошко Анатолий Васильевич,
канд. техн. наук, доцент кафедры электроснабжения Института энергосбережения и энергоменеджмента Национального технического университета Украины «Киевский политехнический институт», Украина, 03056, г. Киев, ул. Борщаговская, 115. E-mail: a-voloshko@yandex.ru

Бедерак Ярослав Семенович,
инженер ПАО «АЗОТ», Украина, 18016, г. Черкассы, ул. Первомайская, 72.
E-mail: yarbak@yandex.ua

Лутчин Тетяна Николаевна,
аспирант Института энергосбережения и энергоменеджмента Национального технического университета Украины «Киевский политехнический институт», Украина, 03056, г. Киев, ул. Борщаговская, 115.
E-mail: t.lutchyn@gmail.com

Кудрицкий Максим Юриевич,
магистрант Института энергосбережения и энергоменеджмента Национального технического университета Украины «Киевский политехнический институт», Украина, 03056, г. Киев, ул. Борщаговская, 115.
E-mail: bugs_bunni@ukr.net

Актуальность работы обусловлена наличием пропущенных данных в показаниях приборов учета энергии.
***Цель работы**: обоснование выбора метода восстановления пропущенных данных об энергопотреблении на промышленных предприятиях.*
***Методы исследования:** модели рассчитываются с помощью приложения Curve Fitting Toolbox программного комплекса «Matlab 7.0». В состав библиотеки графических моделей Curve Fitting Toolbox входит приложение cftool, которое позволяет определить параметрическую модель (например, функции экспоненциальную Exp, полиномиальную Polynomial, рациональную RAT, а также сумму синусоидальных функций SumSin), выполнить подбор параметров, анализ пригодности приближения, отобразить результат графически. В библиотеке графических моделей Curve Fitting Toolbox определяются методом перебора модели из более чем 50 различных математичеких функций.*
***Результаты:** Рассмотрены особенности простых и сложных методов восстановления данных с дальнейшим оцениванием их ошибок (погрешностей). Указаны способы повышения точности n-факторных моделей. Исследованы прямые и обратные зависимости восстановления утерянных учетных данных на примере химического предприятия. Обоснованы оптимальные диапазоны исследования исходных выборок данных. Также предусмотрены варианты определения наиболее рациональных методов восстановления значений в единичных случаях их отсутствия.*

Ключевые слова:
Восстановление данных, n-факторные модели, ошибка модели, энергопотребление, утерянные данные.

Введение

Отсутствие данных технического учета об энергопотреблении на промышленных предприятиях приводит к недоучету энергоресурсов, отсутствию возможности контроля энергоэффективности производственных объектов. Для решения данных во-

просов принято использовать восстановление данных. Восстановление данных необходимо начинать с проверки их выборок на наличие случайных значений.

Знание механизма, приводящего к отсутствию значений, является ключевым при выборе методов анализа и интерпретации результатов [1].

Причинами потери информации об энергопотреблении являются, как правило, следующие [2]: аппаратные и системные отказы, человеческий фактор, программные ошибки, вирусы, кражи и хищения, стихийные бедствия (пожары, наводнения, землетрясения, удар молнии).

Утерянные данные по виду их пропусков принято подразделять на виды [3]:

1) полностью случайные пропуски (data are missing completely at random – MCAR), если условная вероятность не зависит ни от самого пропущенного значения переменной, ни от значений прочих переменных (эта вероятность постоянна для всех наблюдений);
2) случайные пропуски данных (missing at random – MAR), если их вероятность не зависит от самого пропущенного значения переменной, но может зависеть от значений других переменных (в этих случаях механизм пропусков несущественен и к данным применимо большинство методов восстановления пропусков);
3) существенные пропуски данных, если их вероятность зависит от самого пропущенного значения переменной (механизм пропусков является существенным, и для корректного анализа данных необходимо знать этот механизм).

Краткое описание алгоритмов восстановления данных

Можно выделить следующие группы методов заполнения пропусков: простые и сложные [4]. К *простым* (*неитеративным*) *алгоритмам* на основе простых арифметических операций относятся: заполнение пропусков средним арифметическим, метод ближайшего соседа, подбор в группе и регрессионное моделирование пропусков.

Самым простым методом является заполнение средним арифметическим значением по учетным данным. Он не требует применения специального программного обеспечения. Средние значения, вычисленные на исходном и преобразованном массивах, совпадают. Однако такого рода преобразование «усредняет» данные, уменьшая дисперсию признака, и, следовательно, показатели корреляции, что приводит к занижению оценки.

Метод подбора в группе предполагает, что пропуски будут заполнены значениями, полученными в результате оценивания распределения данных по группам. Недостатком данного метода является то, что он требует значительных вычислительных затрат.

При использовании метода парной или многомерной регрессии строится модель линейной зависимости переменной, в которой необходимо заполнить пропуски, от ряда других имеющихся признаков. Регрессионные коэффициенты для каждого из предикторов находятся методом наименьших квадратов в массиве с полными данными. Подставляя значения предикторов в регрессионное уравнение, получают прогноз пропущенного показателя.

Хорошее качество восстановления данных обеспечивает метод сплайн-интерполяции, особенно для одиночных пропусков и небольших выборок. В случае восстановления группы последовательных пропусков результат аппроксимации сплайном данной группы не всегда дает оценки, приближающиеся с достаточной точностью к значениям, которые могли бы быть на месте пропусков [3]. На практике чаще всего используют кубические сплайны и сплайны, не изменяющие форму кривой (сплайны shape-preserving).

Метод экспоненциального сглаживания также применим для восстановления одиночных данных на выборках небольшого объема (например, временной ряд почасовых значений за сутки).

Сложные (*итеративные*) *алгоритмы* предполагают оптимизацию некоторого функционала, отражающего точность расчета подставляемых на место пропуска значений. Их делят на глобальные и локальные.

Особенностью локальных алгоритмов является оценивание (предсказание) каждого пропущенного значения с использованием полного наблюдения, которые находятся в некоторой окрестности предсказываемого объекта.

Глобальные алгоритмы для оценивания каждого пропущенного значения оперируют всеми объектами рассматриваемой выборки. К ним относятся [4]:

- Метод Бартлетта, который представляет собой алгоритм, включающий три итерации. На первой итерации пропуски заполняются некоторым начальным значением (например, средним арифметическим по имеющимся данным). На второй итерации для преобразованной переменной строится регрессионная модель. На заключительном этапе на основе полученного регрессионного уравнения предсказываются новые значения для пропусков.
- Алгоритм Resampling (метод попарного сравнения): выборки данных, содержащие пропущенные данные, заменяют случайно подобранными строками из матрицы полных наблюдений. Затем строится регрессионное уравнение для предсказания отсутствующего значения. Процедура построения регрессионного моделирования повторяется несколько раз. После определенного количества повторений значения полученных регрессионных коэффициентов усредняют и получают окончательное решение с максимальной точностью прогноза пропущенного значения [5].

Особенности n-факторных моделей

Рассмотрим особенности, которые необходимо учитывать при построении моделей восстановления. Чем больше объем исследуемой выборки, тем лучше будут учтены в математической модели особенности ведения технологического процесса. С другой стороны, чем меньше объем выборки, тем меньше влияние сезонных составляющих [6].

Пропуски как зависимых, так и независимых переменных ставят задачу поиска определенного вида математической модели, которую можно использовать для восстановления данных. Для однофакторной модели $y=f(x)$ это может достигаться путем построения моделей вида $y=f(x)$ или $x=f(y)$ с помощью парной регрессии.

Таблица 1. *Однофакторные и многофакторные математические модели для восстановления данных*

Тип математической модели	Метод построения модели	Способы повышения точности построения модели
Однофакторная	Регрессионный метод с использованием пакета CurveFitting-Toolbox и SplineToolbox программы Matlab [7]	1. Метод перебора всех видов математических моделей (экспоненциальных, степенных, полиномиальных, отношения полиномов, суммы синусоид и т. п., в сумме их более 50) [7–11]. 2. Критерии качества моделей: средняя абсолютная процентная ошибка *MAPE*, *F*-критерий Фишера, критерий Акаике *AIC* [12]. 3. Способ определения выборки данных наименьшего объема, что обеспечивает минимальную ошибку модели [12]
Многофакторная	Метод группового учета аргументов [13]	1. Включение в проверочную последовательность характерных точек (например, точки с нулевым энергопотреблением энергоресурсов с координатами (0, 0, 0) при объеме выборки в 150 единиц) [14]. 2. Учет как текущих, так и предыдущих значений независимых переменных [15]. 3. Критерии качества моделей: регулярности $\Delta^2(B)$, минимума смещения, точности кратковременного прогноза $\Delta^2(C)$ и коэффициента простоты $K_{пр}$ [13]

Для двухфакторной зависимости при парном коэффициенте корреляции между зависимой и независимыми переменными более 0,75 для выборок данных за длительный период (300 значений и более) целесообразно строить три однофакторные модели при помощи парной регрессии вместо множественной регрессии [7]. Это проще и эффективнее, чем построение множественных линейных регрессий, которые требуют специального программного обеспечения, подобно методу группового учета аргументов. Результаты исследований в области повышения точности построения однофакторных и многофакторных математических моделей для восстановления данных энергопотребления сведены в табл. 1.

Оценивание ошибок простых и сложных методов восстановления данных

Для построения и сравнения многофакторных моделей рассмотрим данные энергопотребления, расхода ресурсов и выработки продукции на химическом предприятии. Так, если электропотребление E цеха по производству аммиака зависит от объема выпуска аммиака A и от потребления природного газа G, то при наличии пропусков данных и в зависимых, и в независимых переменных и при тесной связи между этими переменными необходимо для восстановления данных строить 3 модели: $E=f(A)$, $A=f(G)$, $G=f(E)$.

Для наглядности фрагмент исходных данных приведен в табл. 2 в виде среднечасовых значений за 01.05.2012.

Таблица 2. *Исходные данные об энергопотреблении*

Время	Выработка аммиака *A*, т	Расход электроэнергии *E*, МВт·ч	Расход природного газа *G*, тыс. м^3
0:00:00	39,699	31,988	45,211
1:00:00	39,292	32,005	45,182
2:00:00	39,644	31,932	45,357
3:00:00	39,929	31,923	45,122
4:00:00	39,684	32,105	45,481
5:00:00	*	32,056	45,782
6:00:00	39,422	32,063	*
7:00:00	43,174	32,098	45,602
8:00:00	42,055	31,971	45,608
9:00:00	40,449	31,953	45,023
10:00:00	41,385	31,860	44,973
11:00:00	38,386	31,759	45,042
12:00:00	39,183	*	45,100
13:00:00	40,331	31,640	44,743
14:00:00	*	31,806	44,836
…	….	…	…

** – пробелы данных.*

Для данных химического производства математические модели определялись тремя простыми методами (замены пропуска средним арифметическим значением, подбора в группе и регрессионным методом) и двумя сложными методами (Барлетта и Resampling), а также методами сплайн-интерполяции кубическим сплайном и одним из методов экстраполяции – методом экспоненциального сглаживания. При восстановлении данных методом Resampling модель строилась без повторений.

Методом парной регрессии рассчитывались ошибки прямых и обратных моделей. Для определения параметрической модели целесообразно использовать отрезки рядов Фурье Fourier, сумму синусоидальных функций SumSin, экспоненциальные Exp, степенные Power, полиномиальные Polynomial, рациональные RAT и другие функции. Далее выполнялся подбор параметров, проводился

анализ пригодности приближения с графическим отображением результата [16, 17]. Затем выбирались лучшие параметрические модели каждого вида функции.

Результаты выбора метода, обеспечивающего наилучшее качество восстановления данных (наименьшую среднюю абсолютную процентную ошибку (*MAPE*)), определяли согласно [18]. Результаты расчетов простых и сложных методов для трехфакторной модели с разными интервалами определения сведены в табл. 3.

Таблица 3. *Результаты расчетов простых и сложных методов для трехфакторной модели с разными интервалами определения*

Количество временных интервалов в выборке	Вид зависимости	Ошибка простых методов, %			Ошибка метода экспоненциального сглаживания, %	Ошибка метода сплайн-интерполяции, %	Ошибка сложных методов, %	
		Метод среднего арифметического	Метод подбора в группе	Регрессионный метод			Метод Resampling	Метод Bartlett
24	$A=f(G)$	0,08	2,42	1,25	2,2	2,42	2,60	1,02
	$E=f(A)$	0,01	0,75	0,26	0,08	0,75	0,10	0,09
	$G=f(E)$	0,01	0,54	0,37	0,11	0,54	0,61	0,24
168	$A=f(G)$	0,91	0,01	2,1	–	–	3,43	9,45
	$E=f(A)$	0,36	0,03	0,53	–	–	0,08	3,62
	$G=f(E)$	0,64	0,03	1,03	–	–	0,06	3,78
700	$A=f(G)$	1,04	0,10	2,07	–	–	2,16	6,53
	$E=f(A)$	0,52	0,01	1,41	–	–	0,91	4,01
	$G=f(E)$	0,62	0,01	1,53	–	–	0,74	2,44
1050	$A=f(G)$	0,94	0,04	1,97	–	–	1,38	2,28
	$E=f(A)$	1,93	0,02	1,26	–	–	0,63	3,91
	$G=f(E)$	0,47	0,02	1,61	–	–	2,56	0,96

После выбора моделей необходимо проверить их результаты на адекватность. Для этого необходимо выбрать степень значимости (например, 0,05) и рассчитать значение F-критерия Фишера, а также F-критическое значение $F_{кр}$. Если $F>F_{кр}$ при данной степени значимости, то модель адекватна согласно работе [18]. Другие критерии, которыми можно оценивать результаты методов восстановления, приведены в работах [19–21].

На основании полученных результатов можно сделать вывод, что лучшее качество восстановления данных энергопотребления в цехе химического производства обеспечивает простой метод подбора в группе. Оптимальным объемом выборки данных, обеспечивающим минимальную ошибку, является минимальная по объему суточная выборка данных (24 значения независимой и 24 значения зависимой переменных). Поэтому исследуемый процесс электропотребления на данном химическом производстве можно отнести к необратимым процессам [22], то есть увеличение числа наблюдений только ухудшит прогнозные и аналитические свойства модели [23].

Как следует из условия стационарности, для наиболее полного анализа стационарных процессов следует собрать как можно больше статистических данных о них. В этом случае удастся тем более точно определить и спрогнозировать характеристики процесса, чем более полной будет выборка наблюдений за ними [22]. Для нестационарных процессов такое правило неприменимо.

Под необратимыми понимаются неоднородные во времени процессы, характеристики которых необратимо меняются с течением времени t и являются вариантными относительно временных сдвигов:

$$t \to t+T, Y(t) \to Y(t+T)+\Delta Y(T)$$

при любом фиксированном T (действительном или целочисленном), где приращение $\Delta Y(T)$ однозначно не вытекает из характеристик процессов в момент времени t. В случае, когда приращения $\Delta Y(T)$ не имеют какой-либо достаточно гладкой тенденции во времени и их изменения непредсказуемы (например, на первом же наблюдении $\Delta Y(T)$ может быть достаточно велико по сравнению с самим показателем $Y(T)$), то такие необратимые процессы хаотические [22].

Хорошее качество восстановления данных для выборки объемом 24 значения показывают методы сплайн-интерполяции и экспоненциального сглаживания. Установлено, что метод экспоненциального сглаживания обеспечивает высокую точность восстановления данных при коэффициенте вариации значений временного ряда до 2 %.

Необходимо указать, что эти методы обеспечивают хорошее качество восстановления одиночных пропущенных данных энергопотребления. При отсутствии нескольких данных подряд лучше использовать регрессионный метод.

Ошибки прямых и обратных зависимостей математических моделей восстановления учетных данных химического предприятия

Рассмотрим способы восстановления данных табл. 2 при помощи различных видов математических моделей (табл. 4). Для расчета утерянных данных использовался регрессионный метод, определяющий ошибки моделей для прямых и обратных зависимостей $E=f(A)$ и $E=f(G)$.

Результаты вычислений табл. 4 указывают на то, что при одинаковых значениях коэффициента парной корреляции значительно отличаются значения ошибок прямых и обратных моделей $E=f(A)$ и $E=f(G)$. Таким образом, один и тот же метод восстановления данных не может обеспечить высокую точность на всем интервале изменения физической величины во времени (на временном ряде). Поэтому для практического применения рекомендуется использовать один из методов, определяющих ошибку модели, используя только временной ряд, и один из методов, учитывающий взаимосвязь между физическими величинами.

Таблица 4. *Способы восстановления данных при помощи различных видов математических моделей для прямых и обратных зависимостей энергопотребления*

Вид математической модели	Объем выборки, ч	$E=f(A)$	$A=f(E)$	Среднее значение	$E=f(G)$	$G=f(E)$	Среднее значение
Exp	24	0,26	2,16	1,21	0,21	0,38	0,29
Fourier		0,27	2,15	1,21	0,30	0,44	0,37
Polynomial		0,28	2,21	1,25	0,21	0,38	0,29
Power		0,28	2,15	1,21	0,21	0,38	0,29
RAT		0,28	2,16	1,22	0,21	0,37	0,29
SumSin		0,27	2,18	1,23	0,31	1,34	0,82
Exp	48	1,13	2,64	1,88	0,95	3,71	2,33
Fourier		3,13	2,69	2,91	1,08	2,99	2,03
Polynomial		1,12	2,79	1,96	1,91	3,73	2,82
Power		1,39	2,69	2,04	2,02	3,74	2,88
RAT		1,11	2,57	1,84	0,97	2,48	1,73
SumSin		0,87	2,57	1,72	0,65	2,06	1,36
Exp	168	0,97	2,26	1,61	0,99	1,55	1,27
Fourier		0,75	6,16	3,45	1,20	1,39	1,29
Polynomial		0,83	1,98	1,41	0,99	1,42	1,21
Power		1,21	2,07	1,64	0,96	1,46	1,21
RAT		0,91	1,91	1,41	0,95	1,22	1,09
SumSin		0,53	1,68	1,11	0,58	1,03	0,80
Exp	350	1,18	1,94	1,56	1,03	1,48	1,26
Fourier		1,14	4,07	2,61	1,10	1,04	1,07
Polynomial		1,14	1,94	1,54	1,03	1,53	1,28
Power		1,31	1,94	1,63	1,03	1,74	1,39
RAT		1,11	1,94	1,52	1,03	1,19	1,11
SumSin		1,24	1,69	1,46	0,85	1,04	0,94
Exp	700	1,66	2,00	1,83	1,37	1,82	1,60
Fourier		1,50	2,08	1,79	1,35	1,70	1,52
Polynomial		1,51	2,13	1,82	1,35	1,71	1,53
Power		1,62	2,09	1,86	1,34	2,14	1,74
RAT		1,47	1,90	1,69	1,32	1,75	1,54
SumSin		1,41	1,85	1,63	1,35	1,53	1,44
Exp	1050	1,46	1,96	1,71	1,71	1,84	1,78
Fourier		1,33	1,95	1,64	1,53	1,67	1,60
Polynomial		1,32	2,12	1,72	1,56	1,71	1,64
Power		1,42	2,08	1,75	1,64	2,07	1,85
RAT		1,36	1,82	1,59	1,61	1,65	1,63
SumSin		1,26	1,83	1,55	1,37	1,61	1,49

Выводы

В результате исследований:

1) упорядочены способы повышения точности построения однофакторных и многофакторных математических моделей;
2) доказано, что лучшее качество восстановления данных энергопотребления обеспечивает метод подбора в группе;
3) сведено восстановление данных сложными глобальными методами Барлетта и Resampling к методу парной регрессии, которая проста в вычислении;
4) предложено для коротких выборок при восстановлении утерянных одиночных данных энергопотребления использовать сплайн-интерполяцию и метод экспоненциального сглаживания, обеспечивающие ошибку моделей до 1...2 %;
5) установлено, что метод экспоненциального сглаживания применим для восстановления данных при коэффициенте вариации значений временного ряда до 2 %;
6) доказано, что процесс энергопотребления химического производства является необратимым процессом;
7) установлено, что ошибки моделей прямой и обратных парных зависимостей не зависят от коэффициента парной корреляции;
8) предложено использовать на практике одновременно один из методов, определящих ошибку модели, используя только временной ряд, и один из методов, учитывающий взаимосвязь между физическими величинами, для обеспечения высокой точности восстановления данных.

СПИСОК ЛИТЕРАТУРЫ

1. Литтл Р.Дж., Рубин Д.Б.А. Статистический анализ данных с пропусками. – М.: Финансы и статистика, 1990. – 336 с.
2. Зинкевич В.С., Штатов Д.К. Информационные риски: анализ и количественная оценка. – М.: Бухгалтерия и бланки. – 2007. – № 1. – С. 50–55.
3. Круглов В.В., Абраменкова И.В. Методы восстановления пропусков в массивах данных // Программные продукты и системы. – 2005. – № 2. URL: http://www.swsys.ru/index.php? page=article&id=528 (дата обращения: 20.01.2014).
4. Бых А.И., Высоцкая Е.В., Рак Л.И. и др. Выбор метода восстановления пропущенных данных для оценки сердечно-сосудистой деятельности подростков // Восточно-Европейский журнал передовых технологий. – 2010. – № 3. – С. 4–7.
5. Злоба Е.А., Яцкив И.Р. Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. – 2004. – V. 6. – С. 51–61.
6. Бедерак Я.С., Лутчин Т.Н., Кудрицкий М.Ю. Влияние объема выборки данных энергопотребления на ошибку математической модели // Международный научно-исследовательский журнал. – 2013. – № 12 (Ч. 1). – С. 37–40.
7. Волошко А.В., Лутчин Т.Н., Бедерак Я.С. Восстановление учетных данных энергопотребления на промышленных предприятиях // Материалы VII МНПК. – Москва, 2012. – С. 179–188.
8. Pentland A., Pentland S. Honest Signals: How They Shape Our World. – Cambridge: MIT Press, 2008. – 208 p.
9. Mayer P. Data Recovery: Choosing the Right Technologies. Datalink, 2003.
10. Holden J.M., Bhagwat S.A., Pat K.Y. Development of a multinutrient data quality evaluation system // J. Food Compos. Anal. – 2002. – № 15 (4). – С. 339–348.
11. Schafer J., Graham J. Missing data: our view of the state of the art // Psychological Methods. – 2002. – № 7 (2). – С. 147–177.

12. Волошко А.В., Бедерак Я.С., Лутчин Т.М. Проблеми вибору оптимальної математичної моделі енергоспоживання на промислових підприємствах // Восточно-европейский журнал передовых технологий. – 2013. – Вип. 5/8 (65). – С. 19–23.
13. Ивахненко А.Г. Долгосрочное прогнозирование и управление сложными системами. – Киев: Техника, 1975. – 312 с.
14. Стеценко І.В., Бедерак Я.С. Побудова багатофакторних математичних моделей енергоспоживання на хімічному виробництві // Энергосбережение, энергетика, энергоаудит. – 2013. – № 7. – С. 41–48.
15. Находов В.Ф., Стеценко І.В., Бедерак Я.С. Застосування методів самоорганізації математичнихмоделей енергоспоживання для встановлення «стандартів» в системах оперативного контролю енергоефективності // Энергосбережение, энергетика, энергоаудит. – 2010. – № 5. – С. 23–33.
16. Дьяконов В., Круглов В. Математические пакеты расширения Matlab. Специальный справочник. – СПб.: Питер, 2001. – 480 с.
17. Відновлення втрачених облікових / А.В. Волошко, Я.С. Бедерак, Т.М. Лутчин, Д.К. Міщенко // Вісник КНУ ім. М. Остроградського. – 2012. – Т. 2 (73). – С. 426–428.
18. Лук'яненко І.Г., Краснікова Л.І. Економетрика: Підручник. – К.: Знання, 1998. – 494 с.
19. Ивахненко А.Г., Мюллер Й.А.К. Самоорганизация прогнозирующих моделей. – К.: Наукова думка, 1985. – 219 с.
20. Горбунов В.М. Теория принятия решений. – Томск: Изд-во ТПУ, 2010. – 67 с.
21. Zwillinger D. CRC Standard Mathematical Tables and Formulae. – Boca Raton: CRC Press, 2003. – 857 с.
22. Светуньков С.Г., Светуньков И.С. Методы социально-экономического прогнозирования. – СПб.: Изд-во СПбГУЭФ, 2009. – Т. I. – 147 с.
23. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985. – 302 с.

Поступила 30.03.2014 г.

UDC 517.4:519.652

THE PROBLEM OF ACCOUNTING DATA RECOVERY ON CHEMICAL ENTERPRISE

Anatoliy V. Voloshko,
Cand. Sc., Institute for Energy Saving and Energy Supply within National Technical University of Ukraine «Kiev Polytechnic Institute»,
115, Borshchagivska Street, Kiev, 03056, Ukraine. E-mail: a-voloshko@yandex.ru

Yaroslav S. Bederak,
AZOT, 72, Pervomayska Str., Cherkassy, 18016, Ukraine.
E-mail: yarbak@yandex.ua

Tetiana M. Lutchyn,
Institute for Energy Saving and Energy Supply within National Technical University of Ukraine «Kiev Polytechnic Institute»,
115, Borshchagivska Street, Kiev, 03056, Ukraine. E-mail: t.lutchyn@gmail.com

Maxim Yu. Kudritskiy,
Institute for Energy Saving and Energy Supply within National Technical University of Ukraine «Kiev Polytechnic Institute»,
115, Borshchagivska Street, Kiev, 03056, Ukraine. E-mail: bugs_bunni@ukr.net

Relevance of the work is caused by the presence of missing data in the readings of energy meters.
The main aim of the research *is to study the choice of method for recovering missing data on energy consumption in industry.*
The methods used in the study: *the models are calculated using the application Curve Fitting Toolbox of the software complex «Matlab 7.0». The library of graphical models Curve Fitting Toolbox includes an application cftool, which allows defining a parametric model (such as, exponential function Exp, polynomial Polynomial, rational RAT, as well as the sum of sinusoidal functions SumSin), selecting parameters, analyzing approach suitability, displaying the result graphically. In the library of graphical models Curve Fitting Toolbox the models from more than 50 different mathematical functions are determined by search method.*
The results: *The paper describes the features of simple and complex data recovery methods with further estimation of their errors and indicates the ways to improve the accuracy of n-factor models. The authors have studied direct and inverse dependences of recovering lost accounting data for a chemical enterprise. The optimal limits of initial research data samples are proved. The paper also provides options for defining the best methods for value recovery in cases of their absence.*

Key words:
Data recovery, n-factor models, model error, energy consumption, lost data.

REFERENCES

1. Littl R.J., Rubin D.B.A. *Statisticheskiy analiz dannykh s propuskami* [Statistical analysis of data with gaps]. Moscow, Finansy i statistika Publ., 1990. 336 p.
2. Zinkevich V.S., Shtatov D.K. Informatsionnye riski: analiz i kolichestvennaya otsenka [Information Risk: analysis and quantification]. Moscow, Bukhgalteriya i blanki Publ., 2007, no. 1, pp. 50–55.
3. Kruglov V.V., Abramenkova I.V. Metody vosstanovleniya propuskov v massivakh dannykh [Recovery Methods omissions in the data]. *Programmnyye produkty i sistemy*, 2005, no. 2. Available at: http://www.swsys.ru/index.php? page=article&id=528 (accessed 20.01.2014).
4. Bykh A.I., Vysotskaya E.V., Rak L.I. Vybor metoda vosstanovleniya propushchennykh dannykh dlya otsenki serdechno-sosudistoy deyatelnosti podrostkov [Selecting a method of reconstructing the missing data for evaluating cardiovascular activity of teenagers]. *Vostochno-Evropeysky zhurnal peredovykh tekhnologiy*, 2010, no. 3, pp. 4–7.
5. Zloba E.A., Yatskiv I.R. Statisticheskiye metody vosstanovleniya propushchennykh dannykh [Statistical methods for recovering missing data]. *Computer Modelling & New Technologies*, 2004, vol. 6, pp. 51–61.
6. Bederak Ya.S., Lutchyn T.M., Kudritskiy M.Yu. Vliyaniye obyema vyborki dannykh energopotrebleniya na oshibku matematicheskoy modeli [Influence of data sampling of energy consumption on mathematical model error]. *Mezhdunarodny nauchno-issledovatelskiy zhurnal*, 2013, no. 12 (P. 1), pp. 37–40.
7. Voloshko A.V., Lutchyn T.M., Bederak Ya.S. Vosstanovleniye uchetnykh dannykh energopotrebleniya na promyshlennykh predpriyatiyakh [Energy recovery measurements in industrial enterprises]. *Materialy VII MNPK* [Materials of VII MNPK]. Moscow, 2012, pp. 179–188.
8. Pentland A., Pentland S. *Honest Signals: How They Shape Our World*. Cambridge, MIT Press, 2008. 208 p.
9. Mayer P. Data Recovery: Choosing the Right Technologies. Datalink, 2003.
10. Holden J.M., Bhagwat S.A., Pat K.Y. Development of a multinutrient data quality evaluation system. *J. Food Compos. Anal.*, 2002, no. 15 (4), pp. 339–348.
11. Schafer J., Graham J. Missing data: our view of the state of the art. *Psychological Methods*, 2002, no. 7 (2), pp. 147–177.
12. Voloshko A.V., Bederak Ya.S., Lutchyn T.M. Problemy viboru optimalnoy matematichnoy modeli energospozhivannya na promislovikh pidpriemstvakh [Problems of choosing the optimal mathematical model of energy consumption in industrial enterprises]. *Vostochno-yevropeyskiy zhurnal peredovykh tekhnologiy*, 2013, vol. 5/8 (65), pp.19–23.
13. Ivakhnenko A.G. *Dolgosrochnoye prognozirovaniye i upravleniye slozhnymi sistemami* [Long-term forecasting and management of complex systems]. Kiev, Tekhnika Publ., 1975. 312 p.
14. Stetsenko I.V., Bederak Ya.S. Pobudova bagatofaktornikh matematichnikh modeley energospozhivannya na khimichnomu virobnitstvi [Construction of multivariate mathematical models of power consumption by the chemical industry]. *Energosberezheniye, energetika, energoaudit*, 2013, no. 7, pp. 41–48.
15. Nakhodov V.F., Stetsenko I.V., Bederak Ya.S. Zastosuvannya metodiv samoorganizatsii matematichnikh modeley energospozhivannya dlya vstanovlennya «standartiv» v sistemakh operativnogo kontrolyu energoefektivnosti [Application of self-organizing power of mathematical models for establishing «standards» in the system of operational control efficiency]. *Energosberezheniye, energetika, energoaudit*, 2010, no. 5, pp. 23–33.
16. Dyakonov V., Kruglov V. *Matematicheskiye pakety rasshireniya Matlab. Spetsialny spravochnik* [Mathematical packets Expansion Matlab. Special Directory]. Saint Petersburg, Piter Publ., 2001. 480 p.
17. Voloshko A.V., Bederak Ya.S., Mishchenko D.K., Lutchyn T.M. Vidnovlennya vtrachenykh oblikovykh [Recovery of lost accounting data]. *Visnyk KNU by M. Ostrogradskogo*, 2012, vol. 2 (73), pp. 426–428.
18. Lukyanenko Í.G., Krasnikova L.I. *Ekonometrika: Pidruchnik* [Econometrics: Textbook]. Kiev, «Znannya», 1998. 494 p.
19. Ivakhnenko A.G., Myuller Y.A.K. *Samoorganizatsiya prognoziruyushchikh modeley* [Self-organization of predictive models]. Kiev, Naukova dumka, 1985. 219 p.
20. Gorbunov V.M. *Teoriya prinyatiya resheniy* [Decision theory]. Tomsk, TPU, 2010. 67 p.
21. Zwillinger D. *CRC Standard Mathematical Tables and Formulae*. Boca Raton, CRC press, 2003. 857 p.
22. Svetunkov S.G., Svetunkov I.S. *Metody sotsialno-ekonomicheskogo prognozirovaniya* [Methods of social and economic forecasting]. Saint Petersburg SPbGUEF Press, 2009, Vol. I, 147 p.
23. Novitskiy P.V., Zograf I.A. *Otsenka pogreshnostey rezultatov izmereniy* [Evaluation of errors in measurement results]. Leningrad, Energoatomizdat Publ., 1985. 302 p.