УДК 519.23 + 519.876.5

УПРАВЛЕНИЕ ПРОЦЕССАМИ НА ОСНОВЕ ДАННЫХ КАК ИНСТРУМЕНТ ЦИФРОВОЙ ТРАНСФОРМАЦИИ ПРЕДПРИЯТИЯ

Дзгоев А.Э.

МИРЭА – Российский технологический университет, г. Москва, Россия, Dzgoev@mirea.ru

Аннотация. На основе информационно-теоретического подхода к моделированию данных показаны вычисления и результаты многомодельного прогнозирования потребления электроэнергии, которое может служить одним из эффективных инструментов цифровой трансформации процесса управления режимами в электроэнергетических системах. На основе данных разработан набор математических моделей-кандидатов для прогнозирования электропотребления. Показано использование информационного критерия Акайке (AIC) и «весов Акайке» для выбора из набора регрессионных моделей-кандидатов наиболее подходящей аппроксимирующей функции, описывающей данные процесса электропотребления. Разработана функциональная модель процесса прогнозирования электропотребления в нотации IDEF0. Сделан вывод о необходимости применения адекватных математических моделей в информационных системах для решения прикладных производственных задач прогнозирования. Отмечена важность применения в учебном процессе моделирования на основе данных.

Ключевые слова. Цифровая трансформация, управление процессами, прогнозирование режимов электропотребления, многомодельный вывод, информационный критерий Акайке (AIC), информационно-теоретический подход, выбор модели.

Цифровая трансформация на предприятии предполагает интеграцию новых решений на основе данных в процессы производства с целью управления ими, для улучшения качества продукции и повышения её конкурентоспособности.

Для возможности управления процессами на предприятии необходимо уметь предвидеть их развитие, с использованием современных способов математического моделирования на основе данных.

При управлении режимами электроэнергетических систем (ЭЭС) одним из инструментов предвидения для принятия управленческих решений является процесс прогнозирования электрической нагрузки. [1].

В настоящее время поток публикаций по тематике прогнозирования электропотребления не уменьшается. Основной побудительной причиной проведения исследований в этом направлении являются высокие и всё ужесточающие требования предприятий к показателям качества прогнозных оценок, например, к точности и достоверности. Так, некоторые предприятия предъявляют высокие требования к ошибке прогнозирования электропотребления, которая должна быть не больше 1%. Вместе с тем, необходимо отметить, что потенциальная экономия электроэнергии всего на 1% позволяет организациям экономить до 7 млн. руб. в год, на примере типового предприятия по производству цинка.

Такие требования заставляют ученых оптимизировать процесс прогнозирования режимов энергопотребления на предприятии путём разработки новых или совершенствованием существующих способов моделирования прогнозирования на основе данных.

В данной статье показана разработанная структура процесса прогнозирования потребления электроэнергии на сутки вперед («to be»), а также результаты проведённого математического моделирования.

В настоящее время всё еще имеются компании, которые составляют прогнозные оценки для предприятий или для региона вручную. Специалисты таких предприятий используют значения электропотребления предыдущих дней, корректируют их, полагаясь на свой опыт. В результате происходят большие ошибки прогнозирования, вследствие которых, в дальнейшем, возникают финансовые затраты.

В данной статье показан один из инструментов цифровой трансформации – моделирование на основе данных.

Рассмотрим контекстную диаграмму процесса прогнозирования электропотребления в нотации IDEF0 [2], которая представлена на рисунке 1.

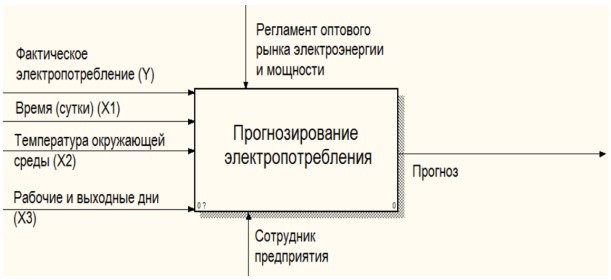

Рисунок 1 – Отображение контекстной диаграммы процесса прогнозирования

Входными данными для процесса прогнозирования являются: фактическое электропотребление за прошлый период; время (сутки); Температура окружающей среды, а также информация о выходных/рабочих днях. Результатом работы процесса является прогнозное значение электропотребления на следующие сутки.

Далее, на рисунке 2, представлена декомпозиция процесса прогнозирования электропотребления, где определены основные подпроцессы участвующие в прогнозировании – первичная обработка данных, моделирование, прогнозирование.

Подпроцесс «Первичная обработка данных» предназначен для выявления аномалий и пропусков в данных, проверок исходных данных на стационарность процесса и исходное вероятностное распределение. Данный подпроцесс особенно важен и может повлиять на качество разрабатываемых математических моделей, так как для получения состоятельных и несмещенных оценок коэффициентов регрессии с помощью метода наименьших квадратов (МНК) не-

обходимо учитывать основные предпосылки регрессионного анализа: 1) Зависимая переменная величина случайная, а независимая переменная – не случайная; 2) Математическое ожидание возмущения равно нулю; 3) Дисперсия возмущений постоянна; 4) Не должно быть автокорреляции в остатках; 5) Зависимая переменная должна быть нормально распределена [3].

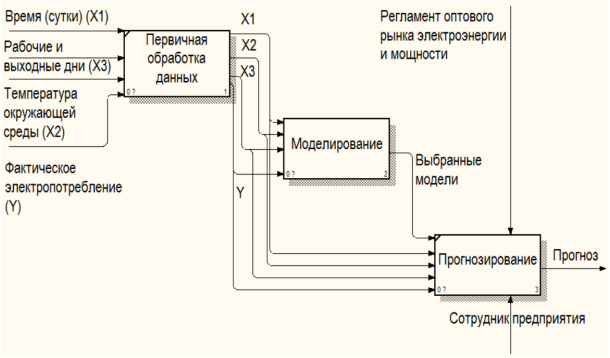

Рисунок 2 – Декомпозиция контекстной диаграммы

Подпроцесс «Моделирование» содержит в себе набор разработанных адекватных моделей-кандидатов и методы расчета коэффициентов для каждого вида моделей. Механизм оценки качества и адекватности каждой модели в наборе (1), а также подпроцесс «Выбор лучшей модели», который использует методы информационно-теоретического подхода. Декомпозиция подпроцесса «Моделирование» представлена на рисунке 3.

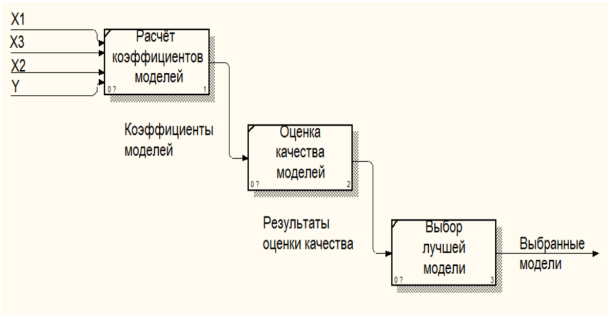

Рисунок 3 – Декомпозиция подпроцесса «Моделирование»

С помощью критериев адекватности и качества необходимо определить модель, которая наилучшим образом описывает процесс электропотребления, сгенерировавший исходные данные.

Выбор лучшей модели является одной из фундаментальных задач научного исследования.

Теоретики информации не верят в понятие истинных (идеальных) моделей. Модели, по определению, являются лишь приближением к неизвестной реальности или истине – не существует истинных моделей, идеально отражающих полную реальность. Британский статистик Джордж Бокс (George Box) , известный своими научными трудами в области планирования эксперимента и анализе временных рядов сделал знаменитое заявление в 1976 году: «All models are wrong but some are useful» («Все модели неверны, но некоторые из них полезны»). Более того, «лучшая модель» на основе данных зависит от размера выборки, так как часто некоторые зависимости могут быть выявлены только при увеличении размера выборки. Например, количество информации в больших наборах данных значительно превышает количество информации в малых наборах данных [4]

Рассмотрим декомпозицию подпроцесса «Выбор лучшей модели», которая представлена на рисунке 4.

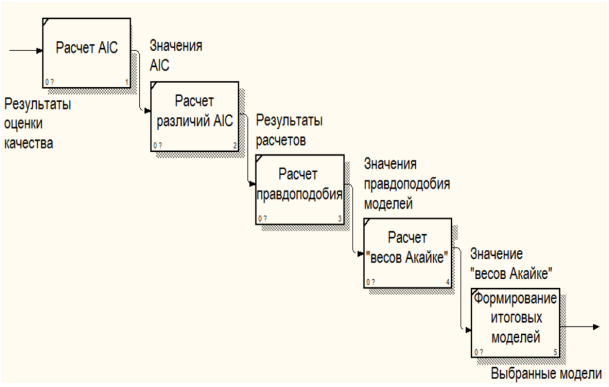

Рисунок 4 – Декомпозиция подпроцесса «Выбор лучшей модели»

Задача, которую решает подпроцесс «Выбор лучшей модели» является достаточно сложной. Выбор подходящей модели из потенциально большого класса моделей-кандидатов – вопрос, который занимает центральное место в задачах регрессии и моделировании временных рядов [5].

Далее в статье рассмотрен один из способов решения задачи выбора лучшей модели, основанный на теоретико-информационном подходе моделирования на основе данных.

Набор моделей-кандидатов

Для моделирования и прогнозирования электропотребления с помощью классического метода наименьших квадратов были разработаны 5 априорных адекватных регрессионных моделей-кандидатов по данным процесса потребления электроэнергии предприятия (количество строк в выборке данных (N) = 25). Структура и вид математических моделей-кандидатов представлены в наборе (1).

$$Y1 = 3.122 \cdot 10^3 + 60.308 \cdot X1 + 5.692 \cdot X2 - \\ -499.251; \_(Модель\_№1\_k = 4).$$

$$Y2 = 7.341 \cdot 10^3 + 105.108 \cdot X1 - 309.164 \cdot X2 - \\ -2.734 \cdot 10^3 \cdot X3 - 3.343 \cdot X1^2 + 5.395 \cdot X1^3 + 0.423 \cdot X1 \cdot X2 + \\ +34.643 \cdot X1 \cdot X3 + 85.978 \cdot X2 \cdot X3; \_(Модель\_№2\_k = 9).$$

$$Y3 = 4.619 \cdot 10^3 + 100.345 \cdot X1 - 66.841 \cdot X2 - \\ -2.442 \cdot 10^3 \cdot X3 - 3.818 \cdot X1^2 + 1.126 \cdot X1 \cdot X2 + \\ +35.666 \cdot X1 \cdot X3 + 72.273 \cdot X2 \cdot X3; \_(Модель\_№3\_k = 8). \quad (1)$$

$$Y4 = 1.426 \cdot 10^4 + 2.832 \cdot X1 - 849.935 \cdot X2 - \\ -3.815 \cdot 10^3 \cdot X3 + 16.482 \cdot X2^2 + 1.107 \cdot X1 \cdot X2 + \\ +35.187 \cdot X1 \cdot X3 + 130.7 \cdot X2 \cdot X3; \_(Модель\_№4\_k = 8).$$

$$Y5 = 4.547 \cdot 10^3 + 123.723 \cdot X1 - 68.245 \cdot X2 - \\ -2.429 \cdot 10^3 \cdot X3 - 6.496 \cdot X1^2 + 0.069 \cdot X1^3 + 1.496 \cdot X1 \cdot X2 + \\ +31.073 \cdot X1 \cdot X3 + 73.948 \cdot X2 \cdot X3; \_(Модель\_№5\_k = 9).$$

где Y1 – Y5 – целевые функции электропотребления, кВт·ч; X1 – независимая переменная, характеризующая время, сутки; X2– независимая переменная – температура окружающей среды, ^{0}C;

X3 – независимая переменная – день недели (1 – рабочие дни, 0 – выходные дни); k – число коэффициентов в модели; N – количество строк в матрице независимых переменных X и матрице-столбце зависимой переменной Y.

Коэффициенты регрессионных моделей-кандидатов были рассчитаны по формуле (2) [6]

$$B = (X^T \cdot X)^{-1} \cdot X^T \cdot Y; \qquad (2)$$

где B – коэффициенты модели; X^T – транспонированная матрица независимых переменных; Y-исходные (фактические) данные электропотребления.

Разработанные статистические аппроксимирующие модели использующиеся для представления процесса потребления электроэнергии, который сгенерировал данные, не могут быть идеально точными. Следовательно, некоторая информация будет потеряна. В этом случае, оценить относительный объём информации, потерянной конкретной моделью можно с помощью информационного критерия Акайке (AIC) (3) [7] который является важнейшим в теоретико-информационном подходе.

$$AIC = N \cdot (\ln(\sigma^2) + 1) + 2 \cdot (k + 1). \qquad (3)$$

где N – количество строк в матрицах данных; σ^2 – дисперсия адекватности, k – число коэффициентов в математической модели.

Таким образом, чем меньше информации теряет модель, тем выше качество этой модели. Учитывая набор моделей-кандидатов, разработанных на основе данных, сгенерированных процессом электропотребления, предпочтительной является модель с минимальным значением AIC.

Результаты оценки адекватности и качества моделей в наборе представлены в таблице 1, где FR – расчетное значение F-критерия Фишера Снедекора, Fт-табличное значение F-статистики, r – значение коэффициента корреляции между зависимой переменной (Y) и расчётным значением зависимой переменной, R^2 – значение коэффициента детерминации.

Таблица 1 – Оценки критериев адекватности и качества по всем моделям-кандидатам в сете

№ модели	AIC	FR > Fт	r	R2	Прогноз кВт·ч.
1	316.27	3.482>2.054	0.865	0.749	4821
2	305.659	6.038>2.235	0.943	0.89	3806
3	304.445	6.23>2.19	0.941	0.886	3780
4	312.665	4.484>2.19	0.918	0.842	4156
5	306.295	5.899>2.235	0.942	0.887	3909

Однако, когда все модели-кандидаты в сете конкурентоспособные, т. е. адекватные и качественные, то определить самую лучшую модель не так легко.

Дело в том, что критерий AIC ничего не говорит об абсолютном качестве модели, а только об **относительном качестве** по сравнению с другими моделями из набора.

Статистический вывод о выборе лучшей модели из набора моделей-кандидатов должен основываться на более чем одной модели, если только данные явно не поддерживают единственную модель описывающую процесс электропотребления [8].

Таким образом, по результатам моделирования в Таблице 1 видно, что у нас недостаточно доказательств о том, что одна из разработанных моделей лучше, чем другие в наборе.

Относительные различия в значениях AIC (Δi)

Для того, чтобы проверить модели-кандидаты на абсолютное качество, необходимо рассчитать остатки значений AIC (Δi) по формуле (4)

$$\Delta_i = AIC_i - AIC_{\min}. \qquad (4)$$

где Δi – относительное различие AIC; AIC_i – значение AIC модели, AIC_{min} – самое минимальное значение AIC в наборе моделей-кандидатов.

Модель, оценённая как лучшая будет иметь $\Delta_i = 0$ или ,другими словами, чем больше значение Δ_i, тем менее вероятно, что подобранная модель является лучшей моделью в информационно-теоретическом подходе Куллбека-Лейблера (K-L) [8, 9].

В 1951 году С. Куллбек и Р. А. Лейблер опубликовали ставшую знаменитой работу [9], в которой авторы количественно определили значение понятия «информация» в связи с концепцией достаточной статистики Р. А. Фишера. Их знаменитый результат, названный *информацией Куллбека-Лейблера* (K-L), является фундаментальной величиной в науке и восходит к концепции энтропии Больцмана. Энтропия Больцмана и связанный с ней второй закон термодинамики представляют собой одно из самых выдающихся достижений науки XIX века [6]. Информацию K-L можно представить как «расстояние» между полной реальностью и моделью.

Итак, нам необходимо выбрать из числа моделей-кандидатов такую модель, которая минимизирует потерю информации. Модель №4 имеет значение AIC = 304.445 – это самое минимальное значение в наборе моделей. Далее, по формуле (4) находим значения Δ_i по каждой модели (5)

$$\begin{aligned}
\Delta_1 &= 316.27 - 304.445 = 11.825_Модель_1 \\
\Delta_2 &= 305.712 - 304.445 = 1.267_Модель_2 \\
\Delta_3 &= 304.445 - 304.445 = 0_Модель_3 \\
\Delta_4 &= 312.665 - 304.445 = 8.22_Модель_4 \\
\Delta_5 &= 306.295 - 304.445 = 1.85_Модель_5
\end{aligned} \qquad (5)$$

Модель №3 имеет значение $\Delta_i = 0$.

В таблице 2 представлены значения Δi и уровни эмпирической поддержки модели.

Таблица 2 – Различия AIC (Δi) и уровни эмпирической поддержки модели [8]

Δ_i	Уровни эмпирической поддержки модели
0-2	Существенный уровень
4-7	Значительно низкий уровень
>10	Нет поддержки модели

Важнейшей особенностью информационно-теоретического подхода является то, что он обеспечивает ранжирование альтернативных моделей, позволяя

сделать некоторые выводы о других моделях, которые также могут быть полезны [8]. Мы можем упорядочить Δi от наименьшего к наибольшему, и такой порядок моделей указывает насколько они хороши в качестве аппроксимации фактически лучшей модели в информационно-теоретическом подходе Куллбека-Лейблера [8, 9]. Ранжирование моделей представлено в таблице 3.

Таблица 3 – Ранжирование моделей по Δ_i

Ранг	Δ_i	№ модели
1	0	3
2	1,267	2
3	1,85	5
4	8,22	4
5	11,825	1

Согласно результатам, представленным в таблице 3, в дальнейшем анализе мы не будем рассматривать априорные модели №1 и № 4 и исключаем их из дальнейшего рассмотрения. Теперь в сете моделей-кандидатов остаётся три модели – №2, №3 и №5, которые имеют близкие значения Δ_i.

Вместе с тем, у нас есть несколько моделей со значениями $\Delta_i < 2$, это означает то, что они сильно конкурируют за позицию лучшей аппроксимирующей модели [10].

Таким образом, у нас есть три сценария: 1) собрать больше данных в надежде на то, что это позволит чётко различать эти три модели; 2) просто сделать вывод о том, что данных недостаточно для поддержки выбора одной модели из первых двух; 3) взять средневзвешенное значение этих трёх моделей с весами, а затем сделать статистический вывод на основе взвешенной мультимодели (многомодельный прогноз).

Хотя различия AIC Δ_i полезны при ранжировании моделей, можно количественно оценить достоверность каждой модели как фактически лучшей модели K-L из моделей-кандидатов. Для этого, необходимо рассчитать Правдоподобие по каждой оставшейся в сете модели (Likelihood of a model) по формуле (6). Правдоподобие показывает относительную силу доказательств по каждой модели.

$$Likelihood = \exp(-\frac{1}{2}\cdot\Delta_i); \quad (6)$$

где Δ_i – различие AIC.

Результаты расчета правдоподобия по оставшимся в наборе моделям №2, №3, №5 представлены в ниже (7):

$$\begin{aligned} &Likelihood\_of\_a\_Model\_№2 = \exp(-\frac{1}{2}\cdot\Delta_i) = 0.531; \\ &Likelihood\_of\_a\_Model\_№3 = \exp(-\frac{1}{2}\cdot\Delta_i) = 1; \\ &Likelihood\_of\_a\_Model\_№5 = \exp(-\frac{1}{2}\cdot\Delta_i) = 0.397; \end{aligned} \quad (7)$$

Чтобы лучше интерпретировать относительное правдоподобие модели с учетом данных и набора моделей-кандидатов необходимо нормализовать значения (7), таким образом, чтобы правдоподобие стало набором положительных «весов Акайке» (w_i) по формуле (8).

$$w_i = \frac{\exp(-\frac{1}{2}\cdot\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\cdot\Delta_r)}; \quad (8)$$

где w_i – значение «веса Акайке»; Δ_r – означает расчет всех моделей в набое.

Значения «весов Акайке» зависят от всего набора моделей-кандидатов, поэтому если модель добавляется или удаляется во время последующего анализа, то wi необходимо пересчитать для все х моделей во вновь определенном наборе.

Результаты вычисления «весов Акайке» преставлены в (9)

$$\begin{aligned} w_2 &= \frac{\exp(-\frac{1}{2}\cdot\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\cdot\Delta_r)} = 0.275; \\ w_3 &= \frac{\exp(-\frac{1}{2}\cdot\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\cdot\Delta_r)} = 0.519; \\ w_5 &= \frac{\exp(-\frac{1}{2}\cdot\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\cdot\Delta_r)} = 0.206; \end{aligned} \quad (9)$$

где w_i – порядковые номера оставшихся моделей в наборе.

«Веса Акайке» рассматриваются как вес доказательств в пользу того, что определенная модель является реально лучшей моделью для рассматриваемой ситуации на основе данных.

«Веса Акайке» обеспечивают эффективный способ масштабирования и интерпретации значений Δi.

Прогнозирование электропотребления на основе нескольких моделей в наборе, используя данные процесса

Если бы одна из моделей в данном исследовании была явно лучшей (то есть $w_i \geq 0.90$), то вероятно, можно было бы сделать прогноз и статистический вывод по одной такой модели. Однако, на практике часто бывает, что ни одна модель в явном виде не превосходит другие в наборе. Очевидно, что возможно вычислить взвешенную оценку прогнозируемой величины с помощью «весов Акайке». Это концепция приводит к модели усредненных оценок (10)

$$\hat{\bar{\theta}} = \sum_{i=1}^{R} w_i \cdot \theta_i; \quad (10)$$

где $\hat{\bar{\theta}}$ – это усреднённая оценка по модели θ_i.

Вышеуказанный способ усреднения моделей полезен для задач прогнозирования.

Прогнозирование – это идеальный способ просмотра усреднения конкурентных моделей, потому что каждая модель в наборе, независимо от ее параметризации, может использоваться для получения прогнозной оценки.

Проведем расчет прогнозной оценки электропотребления по конкурирующим трём моделям (11)

$$Y\_Forecast = \frac{0.275 \cdot 3806 + 0.519 \cdot 3780 + 0.206 \cdot 3909}{0.275 + 0.519 + 0.206} = 3814(кВт \cdot ч.); \quad (11)$$

Фактическое значение электропотребления = 3701 кВт·ч. Следовательно: абсолютная ошибка прогнозирования =| 113 кВт·ч.|; Относительная ошибка прогнозирования = 3,05%.

Внедрение подпроцесса «Моделирование» и техник выбора лучшей модели из сета, позволило получить мультимодальную прогнозную оценку электропотребления и многомодельный статистический вывод, основанный на вычислениях AIC, различий в значениях AIC, правдоподобия и «весов Акайке», которые являются важнейшими критериями для понимания исследуемого процесса, который генерирует данные.

Критерий Акайке относится к информационно-теоретическим методам, которые относительно просты для понимания и практичны для применения в очень большом классе эмпирических ситуаций и научных дисциплин. Однако, информационно теоретические подходы не должны использоваться бездумно; хороший набор априорных моделей является существенным условием, так как это приводит к профессиональному суждению и интеграции науки в набор разработанных моделей.

Информационный критерий Акайке полезен не только для выбора модели для прогнозирования, но также для научного понимания изучаемого процесса, что крайне важно для исследования влияния различных факторов.

Прогнозируется, что «Обратная задача моделирования» (или разработка моделей на основе данных) в ближайшем будущем станет необходимым навыком проектировщиков и разработчиков автоматизированных информационных систем, вследствие появления всё большего количества задач, связанных с прогнозированием процессов на предприятиях, которые невозможно качественно решить без моделирования.

Обратные задачи моделирования используются в системах искусственного интеллекта при разработке моделей. Ценность разработки новых полезных адекватных и качественных математических моделей возрастает в разы и даёт преимущество над конкурентами таким ИТ-компаниям, которые в основе своих систем используют модели. Поэтому, актуально обучать студентов моделированию на основе данных по реальным задачам из практики.

Литература

1. Бэнн Д.В,, Фармер Е.Д. Сравнительные модели прогнозирования электрической нагрузки: Пер. с англ. – М.: Энергоатомиздат, 1987. – 200 с.: ил.

2. Buede, Dennis M. The engineering design of systems: models and methods/Dennis M. Buede. – 2nd ed.p. cm. – (Wiley series in systems engineering and management).

3. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник для студентов вузов, обучающихся по экономическим специальностям / Н.Ш. Кремер. – 3-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2010. – 551 с. – (Серия «Золотой фонд российских учебников») с.439 – 441.

4. Kenneth P. Burnham, David R. Anderson. Multimodel Inference. Understanding AIC and BIC in Model Selection //SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 2, November 2004, 261-304. DOI: 10.1177/0049124104268644

5. Allan D.R. McQuarrie, Chih-Ling Tsai. Regression and Time Series Model Selection. 1998.

6. T. Hastie, R. Tibshirani, J.Friedman. The elements of statistical learning (Data Mining, Inference, and Prediction). 2008.

7. Clifford M.Hurvich, Chih-Ling Tsai. Regression and time series model selection in small samples. Biometrika (1989), 76, 2, pp. 297-307.

8. Burnham, Kenneth P. Model selection and multimodel inference: a practical information-theoretic approach/ Kenneth P. Burnham, David R. Anderson.—2nd ed. 2002.

9. Kullback, S. and Leibler, R.A. On Information and Sufficiency. The Annals of Mathematical Statistics, 22, 79-86. 1951. http://dx.doi.org/10.1214/aoms/1177729694

10. Matthew R.E. Symonds, Adnan Mousalli. A brief guide to model selection, multimodel inference and model averaging in behavioral ecology using Akaike's information criterion.//Behavioral Ecology and Sociobiology. January 2011. Volume 65. Issue 1. Pp.13-21.

DATA-DRIVEN PROCESS MANAGEMENT AS A TOOL FOR DIGITAL TRANSFORMATION OF THE ENTERPRISE

A.E. Dzgoev

MIREA – Russian Technological University, Moscow, Russia, Dzgoev@mirea.ru

Abstract. Based on an information-theoretic approach to data-driven modelling, calculations and results of multi-model forecasting of electricity consumption are shown, which can serve as one of the effective tools for digital transformation of the process of regime management in electric power systems. Based on the data, a set of candidate mathematical models for electricity consumption forecasting has been developed. The use of the Akaike information criterion (AIC) and "Akaike weights" to select from a set of candidate regression models the most appropriate approximating function describing the electricity consumption process data is shown. A functional model of the electricity consumption forecasting process in IDEF0 notation was developed. It is concluded that it is necessary to apply adequate mathematical models in information systems to solve applied production forecasting problems. The importance of applying data-based modelling in the educational process is noted.

Keywords. Digital transformation, process control, power regime prediction, multi-model inference, Akaike information criterion (AIC), information-theoretic approach, model selection.