Factores Climáticos no Consumo de Energia Eléctrica: UM CASO COM UTILIZAÇÃO DE *DATA MINING*

Armando B. Mendes[1], José M.S. Ferreira[2], Aires Ferreira[3]
amendes@uac.pt, 30530001@alunos.uac.pt, airesfer@eda.pt
[1] CEEAplA e Universidade dos Açores, Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal.
[2] Universidade dos Açores, Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal.
[3] Electricidade dos Açores, Rua Dr. Francisco Pereira Ataíde, 9504-535 Ponta Delgada, Portugal.

ABSTRACT:

Este trabalho tem por objectivo identificar causas responsáveis por variações no consumo horário de energia com base na identificação de padrões e relações entre os dados de consumo e várias variáveis climatéricas. Para tal utilizam-se técnicas de *data mining*, nomeadamente a metodologia CRISP-DM e *software* de *data warehouse* MS SQL *Server*. Assim, foi possível verificar que as variáveis climatológicas têm influência muito significativa na produção de energia eléctrica, tendo sido possível prever os consumos de 2007 com um erro absoluto médio de 1,4 MW. Identificam-se ainda vários padrões no comportamento do consumo ou produção de energia eléctrica, nem todos espectáveis face ao conhecimento actual de domínio.

Keywords: data mining, CRISP-DM, consumo de energia, factores climáticos.

CONTEXTO E DEFINIÇÃO DO PROBLEMA

O projecto "efeito de factores climáticos no consumo de energia eléctrica" tem como objectivo analisar e relacionar os registos da produção de energia eléctrica com os registos climáticos na ilha de São Miguel.

Não é claro o efeito dos factores climáticos na variação do consumo de energia eléctrica em determinado contexto. Vários estudos se têm debruçado sobre o difícil problema de prever consumos de energia eléctrica. Como se sabe, não existe nenhuma forma eficiente de armazenar energia eléctrica, pelo que há uma necessidade constante de ajustar a produção ao consumo. Se for possível estimar o consumo com alguma antecedência a produção poderá ser facilmente ajustada, tornando todo o processo mais simples e eficiente. A grande maioria dos trabalhos publicados centra-se na obtenção de previsões o mais precisas possível e com o máximo de antecedência (ver por exemplo: [1], [2], [3] e [4]). Todas estas publicações reconhecem a importância de perceber a relação entre os factores climáticos e o consumo de energia. A previsão dos factores climáticos é a principal fonte de incerteza na previsão do consumo de energia [1].

Este problema é mais complexo em ambientes insulares especialmente sujeitos a alterações climáticas. Assim, para quem tem de dar resposta a essa solicitação de consumo, é necessário caracterizar com cuidado o meio envolvente. As empresas que produzem e/ou distribuem a energia eléctrica têm de tomar decisões em tempo real de quanto produzir por meios flexíveis, principalmente centrais termoeléctricas, de modo a satisfazer a procura em cada segundo.

Neste contexto, qualquer conhecimento adicional sobre a forma como se comporta o consumo de energia eléctrica numa dada região é muito valorizado.

Na ilha de São Miguel a energia eléctrica é obtida principalmente através de centrais termoeléctricas, centrais geotérmicas, centrais hídricas e de sistemas particulares de produção de energia baseados em biogás. Nos últimos anos tem-se verificado um forte aumento da produção de energia proveniente de fontes hidrotermais *energia geotérmica*, tendo atingido no ano de 2008 uma produção média de aproximadamente 41,5% do total da ilha.

Por outro lado, as mesmas empresas mantêm grandes volumes de dados relativos a consumos ou produções horárias de energia eléctrica. Note-se que neste contexto o consumo coincide com a produção uma vez que não se considera a possibilidade de armazenamento. Estes dados podem ser trabalhados de forma a caracterizar o consumo de energia e as necessidades energéticas, a obter informação e conhecimento e, assim, tomar decisões devidamente fundamentadas.

Este trabalho pretende identificar as causas que levam a variações no consumo de energia com base na identificação de padrões e relações entre os dados de consumo e várias variáveis climatéricas. Como meio para alcançar esse objectivo utilizam-se técnicas de *data mining*, para a descoberta de conhecimento, construção de modelos e indução de regras.

Para a concretização desse objectivo, utiliza-se neste projecto a metodologia CRISP-DM *CRoss Industry Standart Process for Data Mining,* já anteriormente descrita pelos autores [5]. A utilização desta metodologia em problemas abordados por técnicas de *data mining* tem-se revelado muito útil por, no essencial, permitirem estruturar e disciplinar o processo, evitando a aplicação indiscriminada de algoritmos como reacção natural à disponibilização dos mesmos em *softwares* de simples utilização.

Note-se, no entanto, que as seis fases do modelo processual nem sempre surgem de forma sequencial, tendo-se verificado mais uma vez a necessidade de voltar um passo atrás com alguma frequência. Estes retornos entre as fases descritas no modelo processual estão previstos na metodologia e constituem a espiral de modelação e extracção de conhecimento [6]. Para uma descrição completa desta metodologia ver o documento original de Chapman *et al.* [7] e o site criado pelo projecto: www.crisp-dm.org.

EXPLORAÇÃO DE DADOS E PRÉ-PROCESSAMENTO

As primeiras dificuldades prenderam-se com a obtenção dos dados e o seu indispensável tratamento para que seja possível a sua utilização com os algoritmos pretendidos.

A base de dados sobre a qual se trabalhou, refere-se aos dados meteorológicos recolhidos no Aeroporto de Ponta Delgada, no período compreendido entre os anos de 1998 e 2007. Fez-se o *download* do site wunderground.com, recorrendo a um programa em Java, construído propositadamente para este fim. Salienta-se que, antes de construir esta ferramenta de trabalho, o *download* dos dados era uma tarefa impensável e quase impraticável, tendo em conta o tempo para a elaboração do projecto, uma vez que era necessário descarregar a informação de 3.300 dias aproximadamente, sendo possível obter apenas um dia de cada vez.

Quanto aos dados dos registos instantâneos das potências de cada um dos sistemas electroprodutores da Ilha de São Miguel, para os anos 2005 e 2006, existem em meio digital de fácil acesso, tornando fácil a importação dos mesmos. No entanto, em relação ao ano de 2007 foi necessário descarregá-lo via intranet da empresa produtora de electricidade na Região Autónoma dos Açores. Devido a limitações no sistema de informação e a restrições de segurança, o *download* destes dados só foi possível em períodos de dez dias. Além disso, estes estavam separados por áreas de produção, dificultando ainda mais a sua recolha.

Além destas dificuldades de acesso aos dados, verificaram-se igualmente alguns problemas durante a exploração dos dados que tiveram por consequência a necessidade de implementar delicadas tarefas de limpeza. Este tipo de tarefas foram efectuadas recorrendo ao SQL *Server* 2005, usando ferramentas OLAP para construir um cubo como ferramenta para resumo e exploração dos dados. O processo de construção do cubo e de implementação de fluxos de processamento (*process flows*) foi semelhante ao descrito em trabalho anterior [5].

Dos problemas mais frequentes, também descritos noutros trabalhos, de que o excelente livro de Chen [8] é um bom exemplo, verificaram-se potências com valores inexistentes ou com valores abaixo do normal e que foram posteriormente considerados como valores omissos (*missing values)*. Muitos problemas semelhantes foram identificados para os atributos relacionados com os factores climáticos.

Finalmente, após ter todos os dados em formato relacional e tendo efectuado as correcção necessárias, fundiram-se numa só tabela todos os atributos referentes ao clima e à potência, sendo assim possível executar um novo cubo com a integração das variáveis climatológicas.

Uma dificuldade relevante nesta fase prende-se com o facto de as chaves nem sempre terem correspondência nos dois conjuntos de dados. Este tipo de problema é conhecido como um problema de fusão de dados (*data fusion*) no sentido em que surge da compatibilização de dados provenientes de fontes diversificadas, ver por exemplo Chen [8] e Saporta [9]. Desta forma, foi necessário confrontar as duas fontes de dados e extrair somente as instâncias cujas horas coincidiam, tal foi efectuado igualmente implementando fluxos processuais adequados.

No final da fase de exploração e pré-processamento, dispunha-se de uma base de dados bem estruturada e com dados no nível de qualidade adequado ao estudo pretendido.

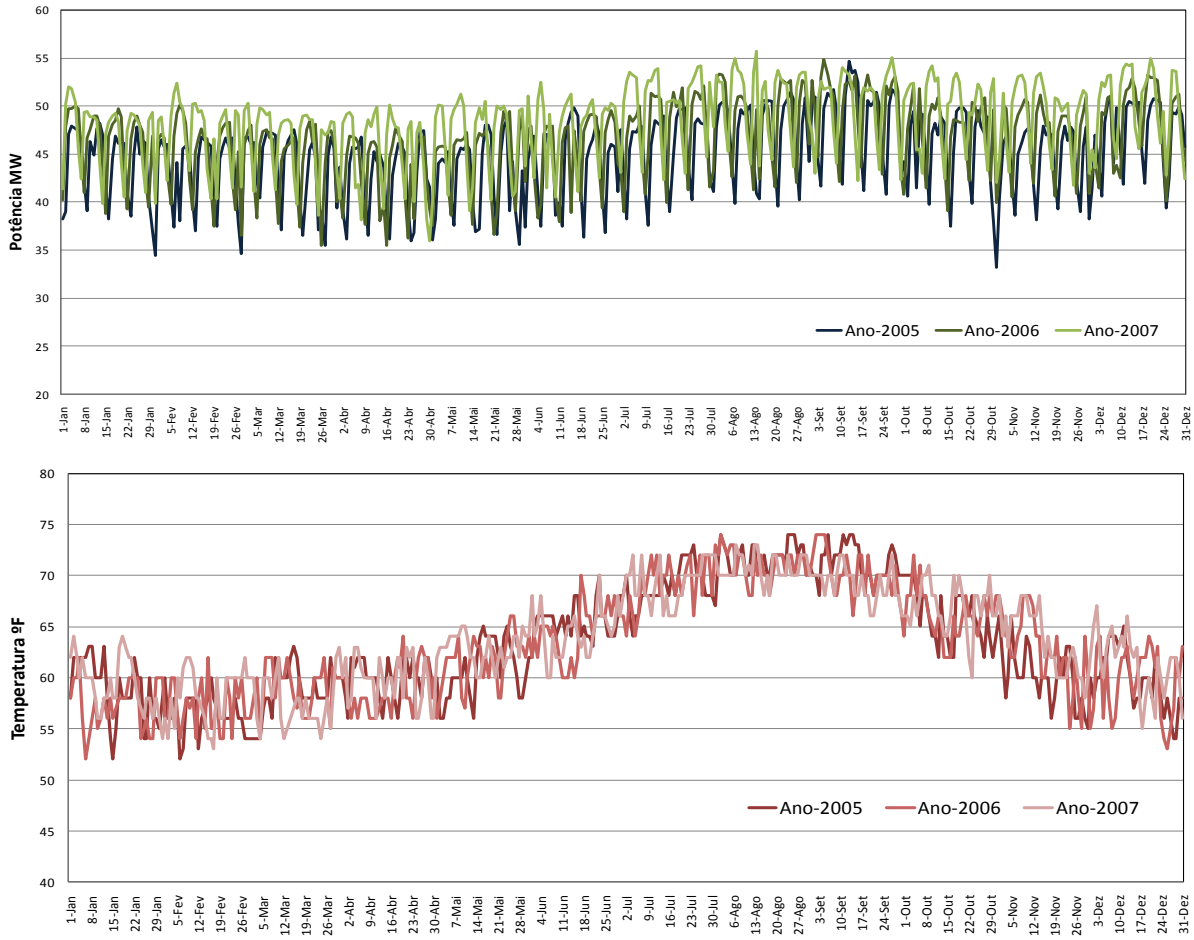

Fig. 1. Três séries de potências médias diárias e as séries correspondentes de temperaturas médias diárias.

CONSTRUÇÃO E VALIDAÇÃO DE MODELOS

Dado o objectivo de identificar relações entre o consumo de energia eléctrica e factores climatéricos, dos algoritmos disponíveis no *MS SQL Server*, os mais adequados para análise explicativa de uma variável quantitativa são o *Microsoft Decision Trees*, *Microsoft Linear Regression* e *Microsoft Neural Network*. Uma breve descrição dos algoritmos disponíveis no software pode ser encontrada em Larson [10].

Na tentativa de encontrar relações entre o clima e o consumo de energia, considerou-se a potência como a variável dependente em todos os modelos. Para compreender o comportamento desta variável ao longo do tempo representaram-se três séries de valores médios diários para três anos na Figura 1. Neste cronograma é possível observar a existência de sazonalidades tanto semanais, apresentando valores menores durante o fim-de-semana, como durante o ano observando-se um aumento do consumo durante os meses de verão e no final do ano. Na mesma figura é igualmente possível observar um aumento anual das potências utilizadas.

No que se refere à temperatura ambiente, o aumento durante os meses de verão é igualmente visível, observando-se uma notável coincidência entre as duas curvas.

Para caracterizar as variáveis utilizadas, começa-se por usar o algoritmo *Naive Bayes*. A divisão em classes ou descretização das variáveis (também conhecido por *bining* não supervisionado) foi realizada pelo mesmo algoritmo e resultou em cinco classes. Desta forma, foi possível verificar, de forma clara, que para temperaturas acima dos 21 ºC as potências situam-se entre os 50 MW e 56 MW, e para valores de temperatura abaixo dos 13 ºC, as potências são menores do que 35,6 MW.

A rede de dependências (*dependency network*), é uma forma de visualização onde os atributos são representados por nós numa rede onde os arcos representam a existência de relações de causalidade (arcos orientados) ou pelo menos algum tipo de dependência entre os nós [10]. Desta análise é possível concluir que as melhores variáveis para prever o consumo de energia eléctrica são, por ordem: a humidade, ponto de

orvalho e temperatura. As piores variáveis são a velocidade do vento e as condições climatéricas.

Por utilização do algoritmo de *Clustering*, utilizando o atributo "Data" como chave, verifica-se que este algoritmo não supervisionado criou um conjunto de dez *clusters*. Alguns desses *clusters* confirmam os resultados anteriores obtidos pelo algoritmo *Naive Bayes*. Por exemplo, num desses grupos a frequência de horas com a potência superior a 47 MW é de 82%, e a frequência da temperatura entre 19,7 e os 27ºC é de 98,3%.

Para a construção de modelos começou-se pela abordagem mais simples estimando um algoritmo de regressão linear múltipla na construção de um modelo cuja variável dependente é a Potência e como possíveis variáveis independentes todas as restantes. O modelo obtido revelou-se de baixa qualidade, não permitindo, nomeadamente, uma boa adaptação a picos de consumo de energia ocorridos durante o dia e que se repetem de segunda a sexta-feira.

Na tentativa de modelar este comportamento, foram elaborados vários modelos de regressão linear. Neste sentido, foram introduzidas variáveis binárias para estimar valores médios de picos de consumo. Adicionou-se ainda uma variável binária que indicasse a existência de fins-de-semana e feriados, uma vez que o comportamento nestes dias é diferente dos restantes dias da semana. Por fim, criou-se uma nova variável "Horas" retirada do campo "Data", que identifica a hora de cada registo.

Na Figura 2 pode-se observar que em parte este objectivo foi, desta forma, conseguido. Contudo, ainda existiam diferenças entre a potência prevista e a realizada com erros significativos e um padrão nos resíduos que não tinha sido modelado. Nesse sentido foram adicionadas novas variáveis binárias que vieram a diminuir os erros nesses pontos.

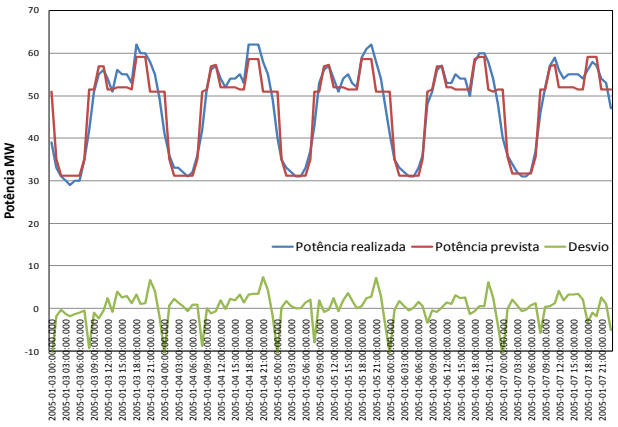

Fig. 2. Valores reais e previstos por rum modelo de Regressão linear, apenas para o mês de Janeiro 2005.

Todo o processo anterior foi repetido, para um espaço temporal mais alargado, utilizando-se registos dos anos 2005 e 2006 num total de 15.666. Esta tabela serviu como dados de treino. Usando todas as variáveis disponíveis construiram-se três modelos com os algoritmos MS *Decision Trees*, MS *Linear Regression* e MS *Neural Network* e verificou-se que aquele que apresenta melhor ajuste aos dados é o que utiliza o algoritmo MS *Decision Trees*.

Na tentativa de melhorar os modelos obtidos introduziram-se igualmente outras variáveis que traduzissem o mês e os dias da semana (variáveis ordinais). Como se verificou na Figura 1, existe sazonalidade nos dados o que justifica ainda a introdução de variáveis binárias identificativas das estações do ano. Para modelar o aumento médio no consumo de ano para ano, criou-se uma nova variável que identificasse o ano.

A opção de introduzir todas estas variáveis binárias, revelou-se acertada uma vez que, ao incluir todos os dias dos anos 2005, 2006, 2007, estas novas variáveis foram escolhidas pelos algoritmos para a construção dos melhores modelos.

Nos novos modelos estimados, os resultados foram muito melhores, obtendo-se os resultados com melhor *score* novamente para o modelo criado pelo algoritmo MS *Decision Trees*. Note-se que este algoritmo, quando são usadas variáveis dependentes numéricas, induz árvores de modelos. Ou seja, é construída uma árvore de decisão de pequena dimensão, de tal forma que em cada nó folha restam um elevado número de registos os quais são utilizados para ajustar um modelo de regressão linear. Obtém-se assim, modelos de regressão para cada segmento de instâncias ou registos.

Durante a modelação, verificou-se com alguma frequência o aparecimento de valores de desvio anómalos, os quais, após algum esforço de pesquisa foram identificados como sendo devidos a valores atípicos das potências em situações de ocorrência de "disparo geral" ou outras semelhantes. Para corrigir estes valores foi necessário recuar à fase de tratamento dos dados. Foi ainda possível concluir que variáveis como a Velocidade do Vento, Direcção do Vento e Condições (identifica acontecimentos climatológicos especiais como tempestades), não são adequadas para explicar o comportamento do consumo de energia eléctrica na ilha de São Miguel.

A validação do melhor modelo obtido com o algoritmo *MS Decision Trees*, foi efectuada com a base de dados de teste correspondente a valores de consumo horário para o ano de 2007. Na Figura 3 apresentam-se algumas estatísticas de

qualidade das previsões obtidas pelo modelo. Da leitura da tabela e por observação do gráfico na figura 4 é possível observar que a qualidade do ajuste é bastante boa.

Coeficiente de determinação múltipla R^2	0,94
Erro quadrado médio (EQM)	3,52
Desvio Padrão (DP)	1,87
Erro Médio (EM)	0,21
Média do erro percentual absoluto (MEPA)	2,92 %
Erro relativo absoluto (ERA)	1,4 %

Fig. 3. Estatísticas de qualidade de ajuste para o melhor modelo obtido.

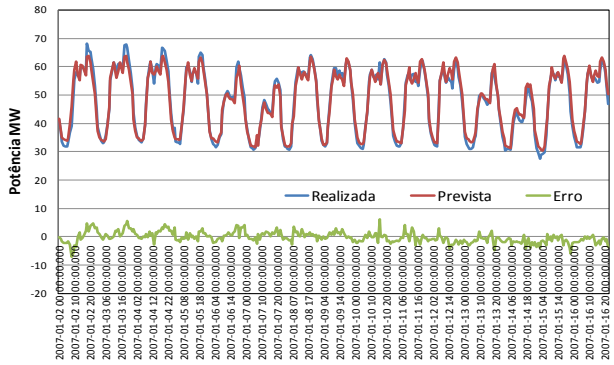

Fig. 4. Potência prevista e realizada para o melhor modelo obtido, ajustado com dados de 2005 e 2006. Dados para o mês de Janeiro de 2007.

RESULTADOS E CONCLUSÕES

Na Figura 5 apresenta-se a expressão do modelo de regressão do nó terminal da ramificação cuja hora está entre as 15 e as 18 às terças-feiras.

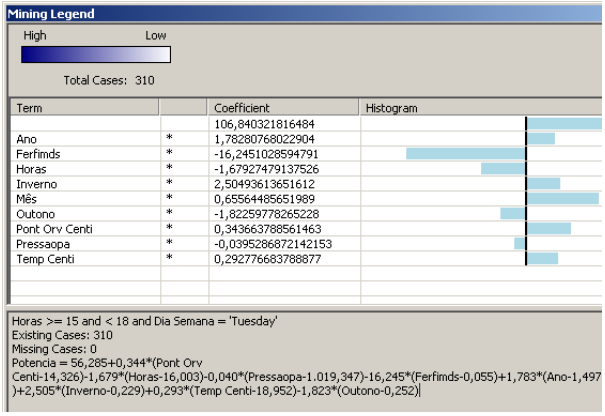

Fig. 5. Um exemplo de um nó folha da árvore de modelos

O modelo obtido pode ser facilmente interpretado. Para todas as terças-feiras entre as 15 e as 18 horas para o aumento de 1 ºC acima da média da temperatura ambiente a potência consumida aumenta 0,293 MW (considerando todas as outras variáveis independentes constantes e com valores médios). Para o mesmo aumento no valor do ponto de orvalho a potência aumenta 0,344 MW. Existe um aumento anual médio de 1,783 MW, um aumento mensal de 0,656 MW, um aumento no Inverno de 2,505*(1-0,229) cujo valor é 1,93 MW, e uma diminuição no Outono de 1,823*(1-0,252) o que perfaz 1,36 MW.

De modo semelhante nas restantes ramificações da árvore de modelos, verifica-se que a temperatura, o ponto de orvalho e a humidade contribuem para a variação do consumo de energia, isto é, quando aumentam os valores destas variáveis há um aumento da potência.

Como se sabe de conhecimento de domínio ou simples senso comum, a existência de elevadas temperaturas obriga-nos a recorrer a sistemas de refrigeração. Por outro lado, quando a humidade relativa do ar é elevada, sabemos que nos sistemas de transporte de energia eléctrica existem perdas devido à corrente de fuga através dos isoladores, verificando-se ainda um aumento da utilização dos desumidificadores por parte dos consumidores.

Ainda assim, a influência de factores climáticos como o ponto de orvalho ou a pressão atmosférica no consumo horário de energia eléctrica não é totalmente evidente do conhecimento de domínio. Considera-se igualmente a possibilidade destas variáveis estarem a funcionar como proxies ou substitutas de factores com influência mais directamente reconhecível no consumo de energia: como a intensidade luminosa do sol ao longo do dia, ou o aumento da população verificado em períodos mais quentes.

Note-se que se considera o modelo de árvores de decisão adequado, não apenas pelos resultados obtidos pelas estatísticas de qualidade de ajuste, mas porque, como se pode observar na Figura 1, as relações com algumas variáveis climatéricas não se mantêm ao longo de todo o ano. Por exemplo, a temperatura pode explicar o comportamento no verão, mas não durante o Natal. Tentar ajustar a mesma regressão a todo o ano resultaria em ajustes de pouca qualidade. Assim, o ajuste por segmentos ou períodos sazonais faz, certamente, mais sentido. Aliás este tipo de comportamento das séries de consumo eléctrico é reconhecido em publicações anteriores sendo recomendada a modelação de períodos específicos como as horas de pico (ver, por exemplo: [3] e [4]).

Fig. 6. Valores médios diários de potência prevista e realizada para o ano de 2007.

A estrutura da árvore de modelos que apresentou os melhores resultados, esquematizada na Fig. 7 é composta por seis níveis. O preenchimento em cada nível é feito segundo o grau de prioridade. No primeiro nível temos como maior prioridade a Hora, seguindo-se os restantes níveis por ordem decrescente de prioridade.

A fim de obtermos uma visão global para o ano de 2007, foram calculadas as médias diárias da potência realizada e da potência prevista como se pode observar na Figura 6. Os valores previstos foram obtidos com os modelos estimados para os anos de 2005 e 2006.

O muito bom ajuste observado confirma a qualidade das previsões. O desvio absoluto médio não ultrapassa 1,4 MW o que significa que esse seria o erro médio da previsão se fosse possível obter valores de previsão das variáveis climatéricas sem erro.

Ao considerar os resultados apresentados pelos modelos é inevitável concluir que o consumo de energia é maior nos períodos de maior temperatura. No Verão a média das potências diárias é mais elevada, embora no Inverno também se verifique um consumo elevado por altura da época natalícia.

Este estudo veio demonstrar que é possível explicar com alguma precisão e exactidão a influência do clima na produção horária de energia eléctrica.

REFERÊNCIAS

[1] Smith, D.G.C. (1989). *Combination of forecasts in electricity demand prediction*. Journal of Forecasting, **8**: p. 349-356.

[2] Troutt, Marvin D.; Mumford, Lloyd G. e Schultz, David E. (1991). *Using spreadsheet simulation to generate a distribution of forecasts for electric power demand*. Journal of the Operational Research Society, **42**: p. 931-939.

[3] Engle, R.F.; Mustafa, C. e Rice, J. (1992). *Modelling peak electricity demand*. Journal of Forecasting, **11**: p. 241-251.

[4] Liu, Lon-Mu e Harris, John L. (1993). *Dynamic structural analysis and forecasting of residential electricity consumption*. International Journal of Forecasting, **9**: p. 437-455.

[5] Mendes, A.B., Ferreira, A. e Alfaro, P.J. (2008). *Suporte à decisão em tecnologias de comunicação: Utilização de OLAP e data mining*, In *Actas CISTI2008,* Cota, M.P., Editor. @LibroTex, p. 973-984.

[6] Lavrač, N., Motoda, H., Fawcett, T., Holte, R., Langley, P. e Adriaans, P. (2004). *Introduction: Lessons learned from data mining applications and collaborative problem solving*. Machine Learning**,** **57**: p. 13-34.

[7] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. e Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide.* SPSS Inc.: USA.

[8] Chen, Z. (2001). *Intelligent Data Warehousing: From data preparation to data mining.* CRC Press: Boca Raton, USA.

[9] Saporta, G. (2002). *Data fusion and data grafting*.Computational Statistics & Data Analysis, **38**: p. 465-473.

[10] Larson, B. (2006). *Delivering Business Intelligence with Microsoft SQL Server 2005.* McGraw-Hill: Emeryville, USA.

Fig. 7. Estrutura lógica da árvore de modelos de regressão completa.

Horas	Dia da Semana	Mês	Temperatura ºC		
[0-3[	≠ Segunda	Mês ≥4,3 e <8,7			
		Mês =4,3	Hora <1		
			Hora ≥1		
		Mês ≥8,7			
	Segunda				
[3-6[	Segunda				
	Terça				
	Quarta				
	Quinta				
	Sexta				
	Sábado				
	Domingo				
[6-9[	Hora <7	Segunda			
		≠Segunda	≠Domingo		
			Domingo		
	Hora >=8	Sábado			
		≠Sábado	≠Segunda	Domingo	
				≠Domingo	
			Segunda		
	Hora =7	Mês <6,5			
		Mês >=6,5			
[9-12[	Segunda				
	Terça				
	Quarta				
	Quinta				
	Sexta				
	Sábado				
	Domingo				
[12-15[	≠Sábado	Hora <14	Mês ≥3,2 e <5,4		
			Mês >=9,8		
			Mês ≥5,4 e <9,8		
			Mês <3,2		
		Hora >=14	Temp. <18,857		
			Temp. >=18,857		
	Sábado				
[15-18[	Segunda				
	Terça				
	Quarta				
	Quinta				
	Sexta				
	Sábado				
	Domingo				
[18-21[	Mês >=10,9				
	Mês >=4,3 e <8,7	Sábado			
		≠Sábado	≠Domingo	Segunda	
				≠Segunda	Mês <6,060
			Domingo		Mês >=6,060
	Mês <4,3	Sábado			
		≠Sábado	Hora <19		
			Hora =19		
			Hora >=20		
	Mês >=8,7 e <10,9				
[21-24[	Mês >=2,1 e <4,3				
	Mês >=8,7 e <10,90				
	Mês >=4,3 e <8,7	Hora >=22			
		Hora <22			
	Mês >=10,9				
	Mês <2,1				