**XX Seminário Nacional de Distribuição de Energia Elétrica**
**SENDI 2012 - 22 a 26 de outubro**
**Rio de Janeiro - RJ - Brasil**

| Rodrigo Bonato Manfredini | Robinson Semolini |
|---|---|
| **ELEKTRO - Eletricidade e ServIços S.A** | **ELEKTRO - Eletricidade e ServIços S.A** |
| rodrigo.manfredini@elektro.com.br | robinson.semolini@elektro.com.br |

# Projeção de Demanda dos Alimentadores da Elektro Utilizando Metodologia de Regressão com Dados em Painel

**Palavras-chave**

Alimentadores
PRODIST
Projeção de Demanda
Regressão com Dados em Painel

**Resumo**

O presente trabalho expõe a metodologia de projeção de demanda por alimentador da Elektro que vem sendo realizada para subsidiar o Planejamento de Expansão do Sistema Elétrico desde o ano de 2010. A metodologia é dividida em duas etapas. A primeira delas é a realização de um algoritmo de clusterização para agrupar os 640 alimentadores da Elektro em grupos homogêneos de acordo com o perfil de consumo e localização geográfica. Na segunda etapa, aplicam-se modelos de regressão com dados em painel nos dados de alimentadores para cada cluster com a finalidade de se projetar a demanda de todos os alimentadores. Os modelos de regressão com dados em painel combinam as metodologias de séries temporais com corte transversal e, portanto, conseguem captar a evolução temporal da demanda máxima bem como características específicas dos alimentadores. Os resultados apresentados por tais modelos nos últimos dois anos tem sido bastante satisfatórios, o que certamente fez com que tal metodologia contribuísse significativamente na melhoria do planejamento de expansão do sistema elétrico da Elektro.

## 1. Introdução

A confecção de modelos estatísticos e econométricos para a realização de projeção de demanda são ferramentas que tem sido cada vez mais utilizadas pelas distribuidoras para subsidiar o planejamento da expansão do sistema elétrico. Os motivos para que tais modelos sejam realizados no modelo atual do sistema elétrico brasileiro são:

- Toda distribuidora contrata anualmente, junto ao ONS, os MUST´s (montantes de uso do sistema de

transmissão) para os próximos quatro anos. Tal contratação deve ser feita da forma mais precisa possível, visto que há penalidades por subcontração ou sobrecontratação e, também, há diversas regras de contratação que exigem grande assertividade das projeções. Por exemplo, não se pode reduzir o MUST total contratado e também não se pode reduzir mais do que 10% o MUST de um determinado ponto. Portanto, os modelos estatísticos realizados devem ter grande qualidade preditiva para que a demanda prevista em cada ponto de contratação seja muito próxima da realizada.

- A cada ciclo de revisão tarifária, as distribuidoras devem calcular o IAS (índice de aproveitamento para subestação). Tal índice é utilizado para avaliação da remuneração dos ativos durante as revisões tarifárias, e considera o fator de utilização da subestação e a expectativa de crescimento da carga da subestação pelos próximos 10 anos. Logo, é necessário que os modelos estatísticos sejam os melhores possíveis para que o planejamento de expansão do sistema seja feito de maneira efetiva, sem excessos de investimentos e nem atrasos na execução de obras. Ademais, uma obra de expansão do sistema pode demorar alguns anos para se concretizar, portanto é extremamente necessário que se tenha uma grande qualidade nas inferências das demandas máximas nos anos subsequentes.

- O Módulo 2 do PRODIST, que trata do Planejamento de Expansão do Sistema de Distribuição, afirma que as distribuidoras devem observar a evolução espacial prevista do mercado para realizar o planejamento da expansão do Sistema de Distribuição de Média Tensão. Então, os modelos estatísticos podem ter grande utilidade para se realizar tais projeções para atender o procedimento.

- A realização precisa de uma obra em uma subestação ou alimentador poderá refletir de maneira significativa no aumento na qualidade de fornecimento de energia, o que acarretará na queda de indicadores de qualidade de fornecimento como DEC, FEC, DIC, FIC e DMIC.

Nesse contexto, o presente trabalho propõe uma metodologia de projeção que pode ter bastante utilidade para se realizar inferências sobre o planejamento da expansão do sistema elétrico. A projeção de demanda por alimentador pode direcionar com grande precisão onde deverão ser feitas as obras de expansão e também pode posicionar quais manobras poderão ser realizadas para uma possível postergação de investimentos no sistema.

O trabalho está dividido da seguinte maneira: Primeiramente será apresentado todo o conjunto de dados utilizado e será discutida a clusterização dos 640 alimentadores da Elektro. A seguir, será apresentada a metodologia de regressão com dados em painel e as justificativas para o uso de tal metodologia. Por fim, serão discutidos os resultados das projeções de demanda por alimentador obtidos.

## 2. Desenvolvimento

A figura 1 mostra um fluxograma de como é realizado o processo de projeção de demanda por alimentador na Elektro. A seguir todas as etapas do fluxograma serão apresentadas e discutidas.

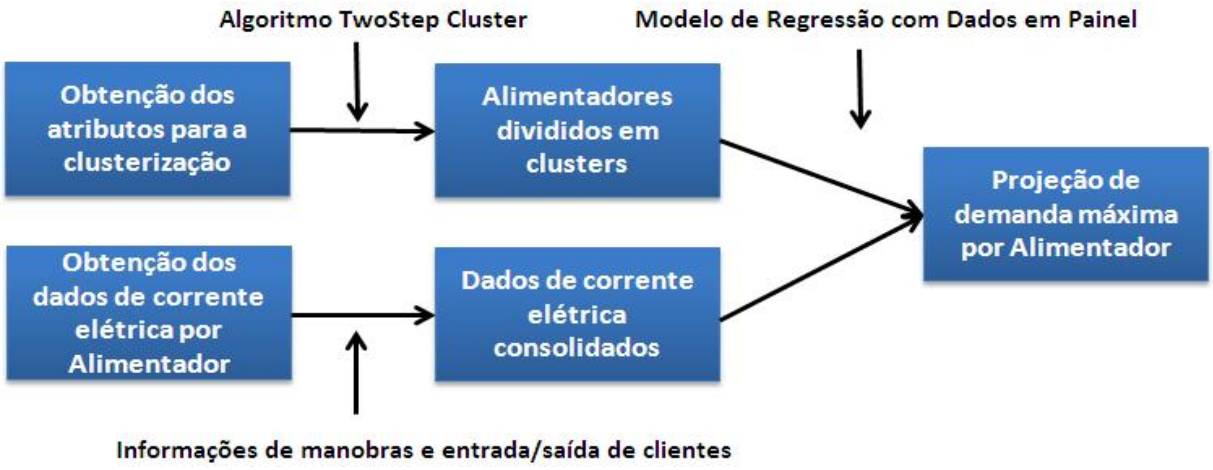


Figura 1: Fluxograma descritivo de como são realizadas as projeções por alimentador na Elektro.

**Obtenção e consolidação das bases de dados**

Para o planejamento de expansão do sistema elétrico da Elektro, ciclo 2011-2016, foram realizadas projeções da demanda máxima de 640 alimentadores. Destes, 487 são de propriedade da Elektro e 153 são de propriedade da CTEEP (Companhia de Transmissão de Energia Elétrica Paulista).

O primeiro passo da metodologia é a obtenção da base de dados de corrente máxima anual para cada um dos alimentadores. É extremamente necessário que o histórico de dados seja a corrente real medida no alimentador, portanto tais dados devem estar isentos de manobras ocorridas e toda entrada e saída de algum cliente relevante no alimentador deve ser analisada.

Adicionalmente, tomam-se todos os dados necessários a nível de alimentador para a realização do algoritmo de clusterização:

- **Região:** Oito regiões distintas utilizadas, conforme mostrado na figura 2;
- **Consumo Total:** Soma do consumo medido anual em 2009;
- **Mix do consumo:** Percentual de consumo residencial, industrial, comercial, rural e outros (poder público, iluminação pública e serviços públicos);
- **Coordenadas X e Y:** Coordenadas X e Y de cada alimentador.


Figura 2: Regiões da Elektro.

## Clusterização dos Alimentadores

O algoritmo de clusterização escolhido para a realização do agrupamento dos alimentadores foi o TwoStep Cluster, utilizando o software SPSS versão 16. Tal algoritmo é bastante rápido quando temos muitas observações para agrupar e também permite a utilização de variáveis categóricas (para maiores detalhes ver MANFREDINI, 2010). Este último ponto é muito importante visto que a variável categórica Região foi utilizada na análise. A distância utilizada na clusterização (métrica para definir se um alimentador/cluster é semelhante ao outro) foi a distância por log-verossimilhança.

Por fim, foram escolhidos dez clusters. Para se verificar a qualidade e coerência da clusterização, faz-se alguns testes de hipóteses e calcula-se também alguns intervalos de confiança. Por exemplo, a figura 3 (a) mostra as estatísticas do teste t para comparação de médias do cluster 1 com a média geral. Verifica-se que, para o cluster 1, a participação do consumo rural é muito mais baixa do que a média geral, pois o valor da estatística t é negativo. Observa-se também que este valor da estatística é o maior em módulo, portanto a variável consumo rural foi a mais importante para a clusterização dentre as variáveis numéricas utilizadas. Ao observarmos a figura 3 (b), verificamos também o fato anteriormente mencionado, pois o intervalo de confiança para a média das participações rurais dos alimentadores deste cluster está abaixo da média geral (linha horizontal na figura 3 (b)) e também está abaixo da maioria dos outros nove clusters. Este é um cluster composto por alimentadores que atendem o litoral da Elektro, cuja característica é um consumo rural quase nulo, baixo consumo industrial e também consumo comercial e residencial acima da média. Para se medir a importância da variável categórica região, fez-se um teste qui-quadrado para cada cluster para se medir a associação entre o cluster e a região. Para todos os dez clusters rejeitou-se a hipótese nula de não associação do cluster com região, mostrando assim que a variável região é bastante importante para a classificação dos alimentadores.

A tabela 1 mostra as características de todos os clusters formados. Observa-se que a escolha de utilizar dez clusters foi bastante razoável, pois o número de clusters por alimentador ficou bem distribuído e os alimentadores pertencentes ao mesmo cluster possuem características semelhantes. A seguir, a metodologia de regressão com dados em painel será discutida e também serão expostos os exemplos de aplicação dos modelos de regressão nos alimentadores divididos em clusters.

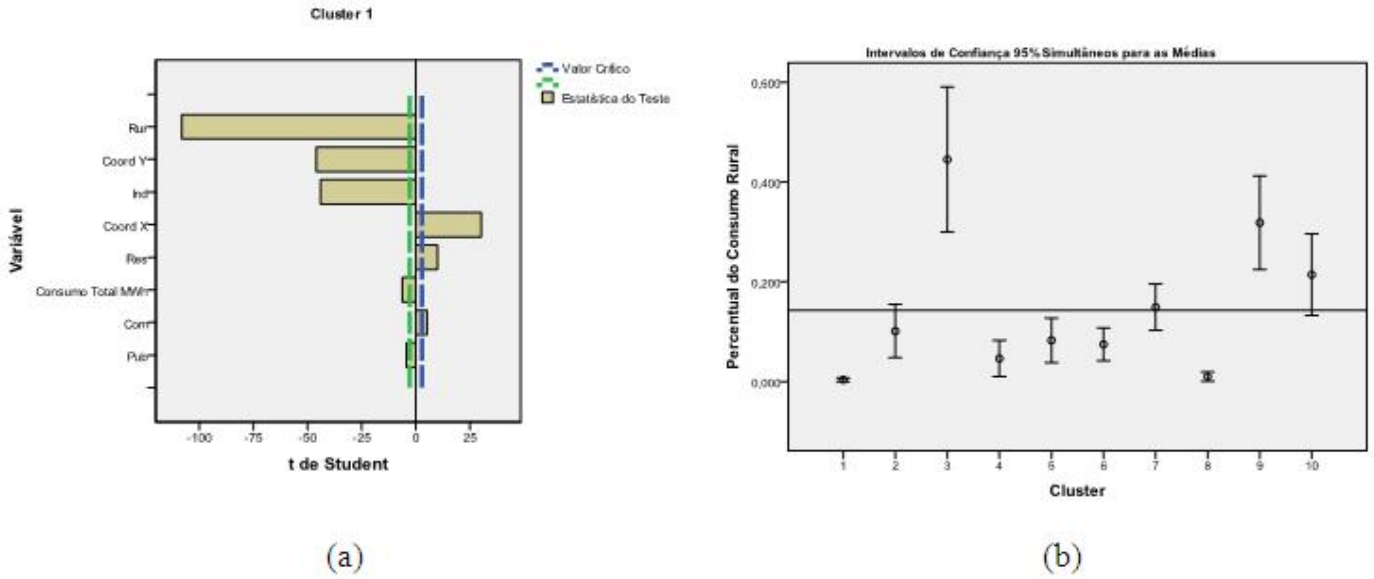


Figura 3: Testes de Hipóteses e Intervalos de Confiança para os Clusters.

Tabela 1: Descrição e Número de Alimentadores de cada cluster.

| Cluster | Descrição | Número de Alimentadores |
|---|---|---|
| 1 | Alimentadores com alta participação residencial e comercial e baixa participação rural. Concentrados no litoral. | 74 |
| 2 | Alimentadores com alto consumo total e participação residencial abaixo da média. Alguns alimentadores da região de Rio Claro e Limeira. | 62 |
| 3 | Baixa participação industrial e alta participação rural. Atendem município a leste do estado de São Paulo. | 42 |
| 4 | Alta participação residencial e baixa participação rural. Atendem municípios na região de Franco da Rocha e Campos do Jordão. | 48 |
| 5 | Alta participação industrial e elevado consumo total. Atendem municípios das regiões de Araras, Tietê e Tatuí. | 66 |
| 6 | Alta participação industrial e elevado consumo total. Baixo consumo rural. Concentrados nas regiões de Limeira, Atibaia e Rio Claro. | 88 |
| 7 | Baixa participação industrial e baixo consumo total. Atendem municípios nas regiões de Peruíbe, Registro, Itapeva, etc. | 79 |
| 8 | Baixa participação residencial, rural e comercial. Alta participação industrial. Atendem municipios na região de Andradina, Três Lagoas, etc. | 36 |
| 9 | Baixa participação industrial e baixo consumo total. Atendem municípios a oeste do estado de São Paulo como Dracena, Pirapozinho, etc. | 85 |
| 10 | Baixa participação industrial. Atendem municípios na região de Ilha Solteira, Jales, etc. | 60 |

## Projeção de Demanda por Alimentador através de Regressão com Dados em Painel

### Regressão com Dados em Painel

Os tipos de dados que, em geral, estão disponíveis para análise empírica, segundo GUJARATI, 2005, são as séries temporais, os cortes transversais e os painéis. As séries temporais são observações de uma ou mais variáveis ao decorrer do tempo. Os dados de corte transversal são dados relativos a uma ou mais variáveis para várias unidades ou entidades amostrais no mesmo período. Nos dados em painel, a mesma unidade de corte transversal, no caso deste estudo, o alimentador, é observado ao longo do tempo. Resumindo, os dados em painel têm uma dimensão espacial e outra temporal.

Há grandes vantagens de se utilizar os dados em painel. Como esta metodologia é uma combinação de série temporal com corte transversal, teremos mais observações para a realização do modelo de regressão e, consequentemente, aumenta-se o número de graus de liberdade. Outro ponto importante é que a regressão com efeitos fixos de entidade com dados em painel controla variáveis não observadas que diferem de uma entidade para outra, mas são constantes ao longo do tempo (STOCK, 2004). Estas variáveis não observadas podem viciar a estimativa dos coeficientes da regressão. Os painéis podem ser classificados em equilibrado e desequilibrado. O primeiro é o caso em que todos os dados são observados no mesmo período em estudo para todas as entidades (no nosso caso, alimentadores). O segundo acontece quando os dados são observados em períodos, com início e/ou término distintos entre as entidades. Os painéis também podem ser estáticos, quando a variável dependente é analisada nos instantes em que são observadas e, portanto, não são consideradas possíveis defasagens da variável na elaboração do modelo. Eles também podem ser dinâmicos, quando a variável dependente é analisada no instante em que é observada e em possíveis defasagens da variável. No caso do estudo em questão, teremos um painel desequilibrado, pois não se tem todos os dados para todos os alimentadores visto que temos alguns alimentadores que entraram em operação depois do ano de 2005 e também temos um painel estático, dado que os melhores modelos escolhidos para cada cluster não utilizaram a variável dependente corrente máxima defasada nos modelos de regressão.

A realização da regressão com dados em painel depende das premissas que são feitas a respeito dos interceptos, dos coeficientes angulares e dos termos de erros. A abordagem dos efeitos fixos consiste na utilização de variáveis binárias do tipo dummy para captar a diferença entre os interceptos das entidades e/ou do tempo. Também se pode utilizar a interação das variáveis dummy a as covariáveis do modelo para captar diferenças entre os coeficientes angulares. A abordagem de efeitos aleatórios assume que o intercepto é uma variável aleatória e as diferenças individuais de cada entidade se refletem em um termo de erro. Na próxima seção, todos os tipos de modelos serão explicados e discutidos.

### Exemplo de Aplicação

As variáveis explicativas utilizadas para a modelagem de regressão com dados em painel foram:

- **Corrente Máxima do Alimentador:** Corrente Anual Máxima em ampere para cada um dos alimentadores no período de 2005 a 2009;
- **Consumo por Alimentador:** Para cada alimentador, tomou-se o consumo residencial, industrial, comercial, rural e outros (poder público, iluminação pública e serviços públicos) em MWh para os anos de 2005 a 2009;
- **Número de Clientes:** Número de clientes atendidos pelos alimentadores de 2005 a 2009;
- **Caged:** Número de empregos formais por município de 2005 a 2009.

Para este artigo, todos os exemplos a seguir serão acerca dos modelos realizados para os alimentadores do

cluster 1. Conforme já dito, tal cluster é caracterizado por conter alimentadores que atendem municípios litorâneos da Elektro como Guarujá, Bertioga, Ubatuba, Itanhaém, Mongaguá entre outros.

O primeiro e o mais simples caso de modelos de regressão com dados em painel é desconsiderar as dimensões de tempo e espaço dos dados combinados e realizar uma regressão habitual estimando os parâmetros por mínimos quadrados ordinários. Ou seja, empilhar as observações dos 74 alimentadores do cluster 1 uma em cima da outra para cada um dos tempos utilizados. Ou seja, temos 74 observações dos alimentadores para cinco anos, totalizando 351 observações (há alguns alimentadores que não temos os dados para alguns anos). O resultado da regressão é o seguinte:

$$Corrente = 151{,}6 + 0{,}015\ Consumo\ Residencial + 0{,}003\ Consumo\ Comercial + \varepsilon \qquad (1)$$

onde $\varepsilon \sim N(0, \sigma^2)$

$R^2 = 0{,}26$, $R^2$ ajustado = 0,25 e Durbin-Watson = 0,55.

Para se chegar no modelo acima, foram testadas todas as variáveis explicativas antes mencionadas e as únicas duas variáveis que foram significativas para explicar a corrente máxima do cluster foram consumo residencial e consumo comercial. Observando-se o modelo acima, vemos que o $R^2$ é muito baixo, ou seja, apenas 26% da proporção total da variação da corrente máxima foi explicada pelas variáveis preditoras. Também tivemos um valor da estatística de Durbin-Watson muito baixo (0,55), indicando que talvez haja autocorrelação nos resíduos do modelo. Se os resíduos do modelo forem autocorrelacionados há problemas ao realizar os testes de hipóteses para verificar se os parâmetros são significativos na regressão linear. Neste modelo, pressupõe-se que o valor do intercepto para os 74 alimentadores é o mesmo (151,6) e também que os coeficientes angulares das duas variáveis dependentes são iguais para todos os alimentadores. A seguir, serão testados alguns modelos com o intercepto e o coeficiente angular variando para cada alimentador.

O segundo caso de modelos de regressão com dados em painel é variar o intercepto para cada alimentador e manter o coeficiente angular constante. O modelo obtido foi o seguinte:

$$Corrente = \beta_{1i} + 0{,}007\ Consumo\ Residencial + 0{,}012\ Consumo\ Comercial + \varepsilon \qquad (2)$$

onde $\varepsilon \sim N(0, \sigma^2)$
i = 1, 2,..., 74

$R^2$ ajustado = 0,68 e Durbin-Watson = 1,56.

O subscrito i no intercepto foi colocado, pois foi estimado um intercepto para cada alimentador. Para se fazer isso, foram colocadas 73 variáveis binárias do tipo dummy. Utiliza-se 73 variáveis binárias ao invés de 74 para não se ter perfeita colinearidade na matriz de desenho e, portanto, a matriz de desenho torna-se inversível.

O modelo (2) parece ser bem melhor que o modelo (1) visto que tem-se um melhor $R^2$ ajustado e também um valor da estatística de Durbin-Watson maior. Também foi realizado um teste F restrito, comparando os dois modelos. Tal teste apontou um valor de F de 7,55 (para 70 graus de liberdade no numerador e 278 no denominador), ou seja, altamente significativo, dando evidências de que o modelo (2) é melhor do que o (1). Pelo modelo (2) pode-se dizer que o aumento de 1 MWh no consumo residencial aumenta em 0,007 amperes a corrente máxima, mantendo o consumo comercial constante. Também pode-se dizer que o aumento de 1 MWh no consumo comercial aumenta em 0,012 amperes a corrente máxima, mantendo o consumo residencial constante. Algumas projeções de alguns alimentadores do cluster 1 realizadas pelo modelo (2) são apresentadas na figura 4.

Uma outra possibilidade de modelos de regressão com dados em painel é a utilização de variáveis binárias tipo dummy para levar em consideração o efeito tempo, ou seja, colocar quatro variáveis binárias e deixar algum ano como referência. Tal modelo apresentou um $R^2$ ajustado de apenas 0,25 e Durbin-Watson de 0,52.

Também há a possibilidade de variar os coeficientes angulares do modelo. Para se fazer isso, deve multiplicar as variáveis binárias dos alimentadores por cada uma das duas variáveis explicativas (consumo residencial e comercial). Tal modelo apresentou um $R^2$ ajustado e Durbin-Watson inferior ao modelo (2).

Todos os modelos acima são da abordagem de efeitos fixos. É preciso se ter em mente que, apesar da metodologia ser bastante eficiente para alguns casos, tal abordagem pode apresentar alguns problemas como excesso de variáveis binárias e consequente diminuição dos graus de liberdade, possibilidade de multicolinearidade dado o elevado número de variáveis no modelo, entre outros.

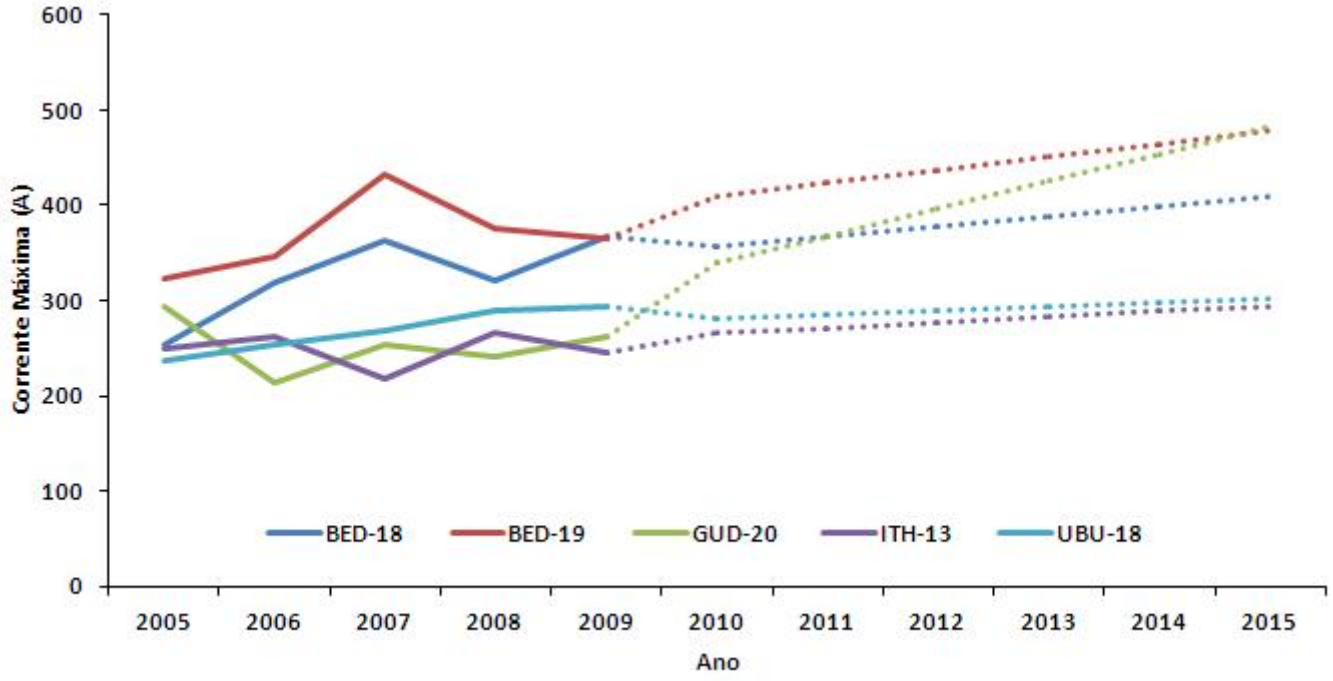


Figura 4: Projeções da Corrente Máxima para alguns alimentadores do cluster 1 utilizando o modelo (2). A linha cheia é o dado realizado enquanto que a linha pontilhada é a projeção.

Uma outra abordagem nos modelos de regressão com dados em painel é a abordagem dos efeitos aleatórios. Ao invés de se colocar variáveis binárias como no caso dos efeitos fixos, tal abordagem expressa o valor do intercepto como uma variável aleatória. Portanto, para o caso do cluster 1, temos:

$$Corrente_{it} = \beta_{1i} + 0{,}013\ Consumo\ Residencial_{it} + 0{,}004\ Consumo\ Comercial_{it} + u_{it} \quad (3)$$

$$\beta_{1i} = \beta_1 + \varepsilon_i\ ,\ u_{it} \sim N(0, \sigma_u^2),\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad (4)$$

i = 1, 2,..., 74
t = 1, 2, 3, 4, 5.

Pensar em $?_{1i}$ como variável aleatória significa que os alimentadores em questão foram tirados de um universo muito maior de alimentadores e eles tem um valor médio comum para o intercepto (que é igual a $?_1$ ) e que as diferenças individuais no intercepto de cada alimentador se refletem no termo de erro $?_i$. Este modelo pode ser vantajoso com relação ao modelo de efeitos fixos devido ao fato de ser mais econômico em graus de liberdade, visto que não se tem que estimar os interceptos individuais. O que se tem que estimar é o valor médio do intercepto e sua variância.

Para o modelo (3) o $R^2$ ajustado foi apenas 0,12 e, considerando o efeito aleatório somente no tempo, o $R^2$ ajustado foi de 0,25. Portanto, para este cluster, o melhor modelo foi o modelo (2). Para cada um dos clusters, todas as abordagens foram testadas e foram escolhidos os melhores modelos de acordo com o $R^2$ ajustado e o valor da estatística de Durbin-Watson.

**Resultados**

Todos os modelos de regressão com dados em painel foram confeccionados utilizando o software Eviews 7. Tal software foi escolhido, pois ele tem uma estrutura específica para se trabalhar com regressão com dados em painel, facilitando, por exemplo, a criação das variáveis dummy para a abordagem de efeitos fixos e também a estimação no caso da abordagem de efeitos aleatórios.

As variáveis explicativas utilizadas em cada um dos modelos são apresentadas na tabela 2. Verifica-se que cada cluster possui suas características específicas, pois as variáveis que explicaram o crescimento da corrente elétrica diferem de um cluster para outro.

Tabela 2: Variáveis explicativas utilizadas em cada modelo de Regressão por Cluster.

| Variável | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 | Cluster 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Caged | | x | | | x | x | | | | x |
| Consumo Comercial | x | x | x | x | | | | | x | x |
| Consumo Industrial | | x | | x | x | x | x | x | x | x |
| Consumo Outros | | x | | x | | | | x | | |
| Consumo Residencial | x | | | x | | x | x | x | x | x |
| Consumo Rural | | x | x | | x | x | | | x | x |
| Número de Clientes | | x | | x | x | | x | | | |

## 3. Conclusões

A metodologia de regressão com dados em painel tem mostrado bons resultados nestes dois anos de aplicação e, portanto, pode ser muito útil para se realizar inferências das demandas máximas por alimentador

para a Elektro. Com a projeção de demanda por alimentador, a distribuidora conseguirá elaborar um CAPEX muito mais preciso para os anos seguintes.

Conforme já dito, a metodologia consiste em uma combinação das metodologias de série temporal com corte transversal, então consegue-se captar a evolução temporal da demanda máxima do alimentador bem como a variabilidade específica de cada um dos alimentadores. Também tem o fato de, ao se trabalhar com dados em painel, se ter uma amostra maior e consequentemente mais graus de liberdade nos modelos. Há também a vantagem de se ter o controle das variáveis não observadas que diferem de uma entidade para outra, mas são constantes ao longo do tempo. O controle destas variáveis pode melhorar significativamente a estimativa dos parâmetros da regressão. Mesmo com tantas vantagens, tem se que ter em mente que há muitos problemas que se pode ter com a metodologia, pois como os dados em painel envolvem dimensões transversais quanto temporais, problemas como heterocedasticidade (variância não constante dos erros) e autocorrelação precisam ser enfrentados. A correlação cruzada das unidades individuais no mesmo ponto de tempo também pode ser um problema. Outro ponto de atenção é a dificuldade de se ter variáveis macroeconômicas na abrangência do alimentador que possam explicar o comportamento da demanda máxima. A única variável macroeconômica utilizada foi o Caged por município (número de empregos formais) devido ao fato de se ter uma atualização constante da variável em nível municipal.

## 4. Referências bibliográficas


CHIU, T., FANG, D., CHEN, J., WANG, Y., JERIS, C. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, p. 263, 2001

GUJARATI, D. *Econometria Básica* Rio de Janeiro, Editora Campus, 4ª Edição, 2005.

MANFREDINI, R., INOUE, M., SEMOLINI, R. Projeção de Demanda de Barramentos Elétricos da Elektro sem Medição Permanente através da Metodologia TwoStep Cluster. *Sendi*, São Paulo, 2010.

STOCK, J., WATSON, M. *Econometria.* São Paulo, Addison Wesley, 2004.

ZHANG, T., Ramakrishnon, R. e Livny, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *Proceedings of the ACM SIGMOD Conoference on Management of Data*, Montreal, p. 103-114, 1996.