ISSN: 2447-5580

DATA MINING DE UMA SMART HOME UTILIZANDO O SOFTWARE R

DATAMINING OF A SMART HOME USING THE R SOFTWARE

Ilson Paulo Peterle Siqueira[1]; Silvia das Dores Rissino[2]

1 Ilson Paulo Peterle Siqueira. Discente do Curso de Engenharia de Computação UFES. Centro Universitário Norte do Espírito Santo - CEUNES. São Mateus, ES. ilsonp@hotmail.com
2 Silvia das Dores Rissino. Docente do Departamento de Computação e Eletrônica DCEL/Centro Universitário Norte do Espírito Santo - CEUNES. São Mateus, ES. silvia.rissino@gmail.com

Recebido em: 08/06/2016 - Aprovado em: 15/07/2016- Disponibilizado em: 15/07/2016

RESUMO: ESTE artigo apresenta a análise de um conjunto de dados de uma smart home (casa inteligente) utilizando o software R. O objetivo do trabalho é realizar uma análise descritiva e preditiva de consumo de energia. Os dados são oriundos do LASS (Laboratory for Advanced Software Systems), que alimentou uma base de dados de uma casa inteligente real. A análise com o software R versão 3.2.4 permitiu a obtenção de um padrão de consumo de energia das áreas da residência.
PALAVRAS-CHAVE: Análise de Dados, Smart Grid, Estatística, Análise Preditiva, Padrão de Consumo.

ABSTRACT: This article presents the analysis of a set of data from a smart home (smart home) using the software R. The objective is to carry out a descriptive and predictive analysis of energy consumption. The data come from the LASS (Laboratory for Advanced Software Systems), which fed a database of a real smart home. The analysis using the R software version 3.2.4 allowed to obtain a pattern of energy consumption residence areas.
KEYWORDS: Data Analysis, Smart Grid, Statistic, Predictive Analytics, Consumption Pattern.

SIQUEIRA, I.P.P. & RISSINO, S.D. (2016). Data mining de uma smart home utilizando o software R. **Brazilian Journal of Production Engineering** (BJPE). 2 (1): 84-98. ISSN: 2447-5580.

INTRODUÇÃO

Smart Grid ou rede inteligente, em termos gerais, é a aplicação de tecnologia da informação para o sistema elétrico de potência (SEP), integrada aos sistemas de comunicação e infraestrutura de rede automatizada. Este tipo de rede envolve a instalação de sensores nas linhas da rede de energia elétrica, o estabelecimento de um sistema de comunicação confiável em duas vias com ampla cobertura com os diversos dispositivos e automação dos ativos (CEMIG, 2016).

As redes elétricas inteligentes surgiram de algumas necessidades causadas pelo constante crescimento econômico do planeta e do padrão de vida das pessoas. O crescimento na demanda de energia elétrica, principalmente em países industrializados, expos a natureza a uma dependência quase insustentável da utilização de combustíveis fósseis. A necessidade de uma rede elétrica mais confiável, necessária pela utilização maior de equipamentos digitais mais sensíveis a variações de carga, é cada vez mais necessária. Em contraponto, a rede elétrica existente é antiga e tem a necessidade de interligação com novas fontes de energia.

O smart grid aparece então como um conjunto de soluções para melhorar todo sistema elétrico, desde a geração até a distribuição, integrando informações da rede, entre concessionárias e clientes, e permitindo a integração de fontes de energia renováveis.

O smart grid, seus medidores e tecnologias inteligentes estão transformando a forma como a rede elétrica funciona, desde a geração passando pela transmissão até a distribuição de energia. Esta transformação permite que a experiência dos geradores, distribuidores e dos consumidores evolua para a busca de novas formas de energia que suportam sistemas mais colaborativos, ecológicos, estáveis, confiáveis e com um custo mais eficiente no geral (FRYE, 2008).

Novas tecnologias inseridas no smart grid mudam completamente o volume de informações coletadas da rede elétrica, tanto dos consumidores quanto dos geradores e distribuidores. Para lidar com esse grande volume de dados, gerados por toda a tecnologia, faz-se necessário avaliar sistemas que trabalhem com armazenagem, mineração e análise (AUNG et al, 2012; MACK, 2014).

Este artigo propõe o uso da linguagem R devido a possiblidades de trabalhar com grandes volumes de dados, aos benefícios de possuir vários pacotes para análises estatísticas e, aos muitos exemplos de usos em várias áreas possibilitando fácil aprendizado de uso. Através da utilização desta ferramenta foram apresentadas análises tanto descritivas quanto preditivas dos dados, utilizando modelos estatísticos. O objetivo foi, através da análise de dados, fornecer *"insights"* para o consumidor, permitindo demonstrar o perfil de consumo e estado de sua rede, bem como fornecer previsões de consumo da residência (HOSSAIN; AMANULLAH; ALI, 2010).

SMART GRID

Smart Grid pode ser descrito como uma rede transparente e sem costuras, que "entrega" informações de energia em dois sentidos e, permite a indústria de energia gerenciar melhor a distribuição e transmissão, possibilitando também aos consumidores mais controle sobre as decisões do setor energético (HOUSSAIN,2010; AMANULLAH, 2010; ALI, 2010).

O objetivo do Smart Grid é proporcionar uma melhor visibilidade da rede elétrica bem como permitir um envolvimento maior dos consumidores na função de geradores, através de medidores inteligentes (smart meter) e casas inteligentes (smart home). O smart grid incorpora atributos avançados de tecnologias de informação e comunicação para conduzir informação em tempo real e facilitar a estabilidade instantânea do fornecimento e demanda da rede elétrica. Os dados operacionais coletados pelo smart grid e seus subsistemas permitem aos operadores do sistema reconhecer as melhores linhas de ação para proteger a rede contra instabilidades e incidentes (HOSSAIN; AMANULLAH; ALI, 2010).

> O Smart Grid usa tecnologias digitais para melhorar a confiança, segurança e eficiência (econômica e energética) do sistema elétrico desde a geração, através do sistema de transmissão até a distribuição (consumo) e um através de um grande número de sistema de geração distribuídos e sistema de armazenamento. (U.S. Department of Energy, 2009, p.1).

Para cumprir as necessidades dessa "nova" rede elétrica várias tecnologias deverão ser desenvolvidas: adaptações na rede elétrica existente, instalação de sensores, emprego de novos equipamentos, o desenvolvimento de uma rede de comunicação para suportar essas transferências de dados entre os pontos da rede elétrica.

Nesse caso, para suportar os dados e informações gerados por essas tecnologias, serão necessárias implementações de tecnologias para armazenamento e tratamento de dados. Dada a concepção de monitoramento em tempo real dessa rede, essas tecnologias deverão permitir o tratamento de grande volume de dados. Sistemas de banco de dados deverão conseguir trabalhar com grandes volumes de dados e em tempo certo (tempo real ou quase tempo real), poderá ser necessário trabalhar com paradigmas diferentes da usual como banco de dados relacionais(AUNG et al, 2012).

Também serão necessárias ferramentas e modelos de análises dessa grande quantidade de dados, empregando softwares para mineração de dados, análise e consumo de informações. Estas ferramentas devem possibilitar a manipulação de grande volume de dados, extraindo conhecimentos da rede e permitindo previsões, baseadas em modelos. Isto permitirá acompanhamento da estabilidade da rede, consumo e demais informações. Além disso, permitirá o cruzamento de informações de demanda e geração, não só das fontes convencionais como dos novos sistemas que poderão ser inseridos na rede, possibilitando uma previsão mais acurada do uso de energia elétrica. Essas informações poderão ser úteis tanto para a indústria de energia, provendo seus gestores com informações para tomada de decisão, como também para os consumidores que, em suas casas inteligentes, poderão acompanhar melhor o seu perfil de consumo diário e fazer previsões, podendo até fazer campanhas de economia de energia pessoais, ou quando necessário por limitações de fornecimento (FREY, 2008; HOSSAIN; AMANULLAH; ALI, 2010; EKANAYAKE; LIYANAGE; WU, 2012).

DADOS DE UMA SMART HOME

Uma smart home (casa inteligente), é uma residência dotada com equipamentos que permitem além do controle de dispositivos eletrônicos, o monitoramento de informações como consumo de energia, status de equipamentos, temperatura e outros dados (IoT Agenda, 2016).

Uma Smart Home, na visão do smart grid, é dotada de equipamentos que permitem o acompanhamento do seu consumo elétrico, geração de energia. Esse acompanhamento é possível pelas instalações de smart meters (medidores inteligentes) que fornecem informações do consumo de energia. Tais equipamentos podem ser instalados em ramais elétricos, o que dará informações mais precisas sobre o consumo dos equipamentos.

Para o propósito deste artigo foi utilizado um dataset (conjunto de dados) fornecido por Laboratory for Advanced Software Systems (LASS, 2016). O conjunto de dados compreende uma grande variedade de informações de uma casa real, incluindo consumo e geração elétrica, monitoramento do ambiente (temperatura, umidade etc.), e eventos operacionais de interruptores, por exemplo.

A coleta foi realizada compreendendo, em sua maioria, um período de um mês com frequência de alguns minutos, podendo variar de acordo com a origem da informação e do equipamento.

As informações estão agrupadas em pastas, por tipos de dados coletados, em cada grupo existem arquivos no formato .csv, representando cada dia de medição. Um arquivo FORMAT também foi fornecido para indicar o que cada coluna dos arquivos .csv representa.

O arquivo do tipo .csv (comma-separated values) é um formato que possui os dados separados por algum delimitador, que pode ser uma vírgula, tabulação, ponto e vírgula etc.

O SOFTWARE R

A escolha do software para realizar as análises descritivas e preditivas levou em consideração a necessidade de se trabalhar um grande volume de informações, permitir que as análises necessárias fossem realizadas, ou seja, possuir funções estatísticas e ainda dar suporte à aplicação de modelos matemáticos, e também pelo fato de ser open source.

O software R atende às necessidades e possui ainda uma grande base de conhecimento na internet. O R é uma linguagem e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos. Além disso é um ambiente altamente expansível, pois além de pacotes disponíveis com sua instalação, permite a inclusão de muitos outros pacotes disponíveis em sua rede de distribuição CRAN

(Comprehensive R Archive Network) (R Foundation, 2016).

A linguagem escolhida se adapta então as necessidades propostas para o trabalho, disponibilizando ampla variedade de técnicas estatísticas e gráficas, aplicação de modelos, além de possuir várias interfaces gráficas. A interface gráfica escolhida para o trabalho foi o R Studio.

TRATAMENTO DE DADOS

A base de dados escolhidas para o trabalho forneceu grande variedade de informações. Para exemplificar o uso das ferramentas serão utilizados os dados de consumo dos circuitos de uma casa. Estes circuitos estão divididos geralmente por grupos, cômodos ou equipamentos, por exemplo, o consumo das luzes da sala de estar, ou o consumo do forno micro-ondas.

DISPOSIÇÃO DOS DADOS

Os dados dos circuitos foram dispostos em um arquivo csv separado por virgulas como apresentado em Figura 1.

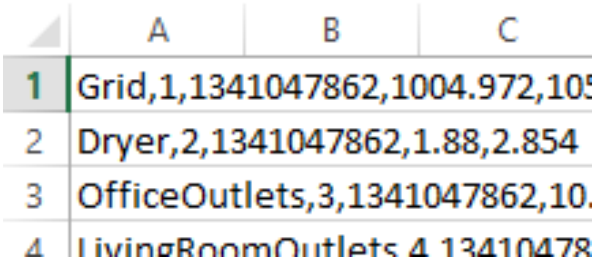

Figura 1 - Disposição dos dados no arquivo csv.

Os dados representam potências ativas medidas de cada circuito da casa durante o dia. Para verificar o que cada coluna de dados representa deve-se analisar o arquivo FORMAT fornecido junto com os dados.

Apesar da limitação dos dados para o trabalho de mineração de análise, o dataset escolhido possui outras informações que podem enriquecer as análises. Uma evolução do trabalho analítico pode ser o cruzamento de dados de consumo com dados de controle dos equipamentos da casa, ou dados do meio ambiente, como temperatura e outros fatores climáticos com dados de geração de energia eólica e solar. E esse dataset é suprido com todas essas informações.

CARREGAMENTO DE DADOS NO R

Para fazer a carga dos dados para o software R foi necessário um conhecimento prévio das informações a serem trabalhadas. Para o trabalho foram escolhidos os dados de consumo dos circuitos da casa. Para cada dia de medição dos circuitos existe um arquivo com as informações de todos os circuitos. Essas medições são: nome do circuito, número do circuito, timestamp (cadeia de caracteres denotando a data ou hora que um evento ocorreu), potência real do circuito, potência aparente do circuito. Os títulos de cada campo estão em um arquivo separado, FORMAT.

O tratamento inicial dado é o carregamento dos dados para o R, é feito através de funções básicas de carregamento de arquivos csv para trabalhar com a base proposta. O carregamento de dados é realizado com funções diferentes a depender do formato de origem dos dados, uma pesquisa sobre as funções de carregamento foi necessária para conhecer quais funções utilizar.

Nos dados tratados nesta análise utilizaremos a função read.csv, que lê os arquivos com este formato, com uma pequena modificação para leitura em lote de vários arquivos. O script 1 mostra como realizar a leitura dos dados de um arquivo.csv para o Software R.

Script 1 – Realiza a leitura dos dados de um arquivo .csv para o R.

```
basecircuitos<-read.csv(arquivo,  header  = FALSE, sep = ",")
```

A função acima carrega os dados de arquivo csv, considerando que o arquivo não possui os cabeçalhos das colunas, e que é separado por vírgulas. As informações serão armazenadas em uma entidade do R dataframe chamada basecircuitos.

Ao lidar com vários arquivos representando dias de um mês, a utilização de uma função para ler arquivos em lote agilizou o trabalho. No código seguinte é mostrada com criar uma função chamada load_data que realizará a leitura de vários arquivos csv que estiverem em um diretório. O script 2 faz a leitura dos dados de vários arquivos .csv para o R.

Script 2 – Realiza a leitura dos dados de vários arquivos .csv para o R.

```
load_data<-function(path) {
   files <- dir(path, pattern = '\\.csv', full.names = TRUE)
   tables <- lapply(files, read.csv,header=FALSE)
do.call(rbind, tables);}
```

A função acima pode ser declarada no R e logo após ser executada com o caminho (diretório) dos arquivos a serem lidos. Todos os arquivos serão lidos e, portanto, deve ser tomado o cuidado para que o diretório possua somente os arquivos .csv a serem lidos.

Após o carregamento dos dados foi criado um dataframe chamado basecircuitos que possui os dados de forma “bruta” assim como mostrado na figura 2. Dataframe é a entidade no R utilizada para guardar as informações, as tabelas. É na verdade uma lista de vetores de tamanho igual.

Filter

	V1	V2	V3
1	Grid	1	1341047862
2	Dryer	2	1341047862
3	OfficeOutlets	3	1341047862
4	LivingRoomOutlets	4	1341047862

Figura 2. Dados carregados no R.

Verificando a Figura 1 e Figura 2 é possível fazer uma correlação dos dados em csv e dos dados carregados no R.

A partir daí, para uma melhor identificação, foi necessário renomear as colunas, que representam os campos, conforme descrito no arquivo FORMAT.

A renomeação das colunas foi feita manualmente utilizando a função colnames mostrado no script 3, ou fazendo o carregamento do arquivo FORMAT, neste caso utilizando a função read.table, e logo após atribuindo os dados do dataframe gerado deste carregamento para os nomes das colunas da basecircuitos. O resultado é apresentado na Figura 3.

Script 3 – Realiza a renomeação dos nomes das colunas do dataframe.

```
Colnames(basecircuitos)<-
c(“CircuitName”,“CircuitNumber”,
”TimestampUTC”,”RealPowerWatts”
,”ApparentPowerVAs”)
```

Filter

CircuitName	CircuitNumber	TimestampU
Grid	1	13410
Dryer	2	13410
OfficeOutlets	3	13410

Figura 3. Dados com colunas já renomeadas.

Após o carregamento e identificação dos dados foi necessário fazer alguns tratamentos. Inicialmente a verificação dos tipos e formatos de dados.

TRANSFORMAÇÃO DE DADOS

O dataframe basecircuitos possui agora os dados de potência dos circuitos da casa. Esses dados são representados por colunas numéricas, texto (caractere) e data-hora. Foram necessárias transformações na coluna que possui o nome do circuito, para que seja representada como texto, e na coluna da data hora, para que represente a data e hora no formato local.

A identificação da adequação de tipos e formatos é inerente aos dados e ao tipo de tratamento pretendido. No dataframe criado, transformações necessárias ao trabalho serão realizadas.

Através da função mostrada no script 4 pôde-se identificar os tipos de dados, recebendo a seguinte resposta apresentada na Figura 4.

Script 4 – Exibe o tipo de dados do dataframe basecircuitos.

```
sapply(basecircuitos,typeof)
```

```
CircuitName     CircuitNumber        Ti
  "integer"         "integer"
```

Figura 4. Tipos de dados do dataframe basecircuitos.

Os campos (colunas) CircuitName e TimeStampUTC foram transformados em tipos caractere e data-hora, respectivamente.

Primeiramente a transformação do campo CircuitName pode ser facilmente feita utilizando o seguinte comando do script 5:

Script 5 – Realiza a transformação de tipo de uma coluna para texto.

```
basecircuitos$CircuitName<-
as.character(basecircuitos$CircuitName)
```

O comando foi executado para dizer ao R que transforme os dados do campo desejado para o tipo caractere (texto).

O dado de timestamp, com a data e hora da medição, demandou um pouco mais de trabalho para correção. Na base analisada, a timestamp está formatada no padrão UNIX. Este formato é simplesmente um contador com início em 01/01/1970 00:00:00. Para fazer a transformação deste dado, que originalmente está numérico, foi utilizada a função do script 6:

Script 6 – Realiza a transformação de uma coluna do tipo timestamp para o formato de data desejado.

```
basecircuito$timestamp<-
as.POSIXct(basecircuito$timestamp,
origin="1970-01-01")
```

Neste caso o campo (coluna) timestamp recebeu seu próprio valor transformado do formato timestamp unix, para o formato necessário “ano-mês-dia horas:minutos:segundos”.

Essas são duas transformações básicas de tipo de dados necessárias para preparar os dados para análises posteriores. Estas foram as transformações aplicadas, ao tratar outros tipos de dados modificações diferentes, em outros trabalhos, podem ser necessárias. O R fornece uma gama de funções que tratam dessas transformações sem muitos problemas. Quaisquer transformações adicionais necessárias podem facilmente ser programadas na linguagem R.

Ao verificar novamente os tipos de dados foi identificada a transformação da coluna CircuitName e TimestampUTC, como observado na Figura 5. Também é visível o novo formato apresentado para TimestampUTC na Figura 6.

```
CircuitName     CircuitNumber        Times
 "character"       "integer"
```

Figura 5– Tipos de dados transformados.

	CircuitNumber	TimestampUTC	Re
	1	2012-06-30 06:17:42	
	2	2012-06-30 06:17:42	
	3	2012-06-30 06:17:42	
	4	2012-06-30 06:17:42	

Figura 6 – Novo formato de TimestampUTC.

Após todas as transformações foi possível verificar que o dataframe possui a potência dos circuitos, real e aparente por circuitos da casa, esses dados foram

coletados em intervalos bem pequenos, as vezes com variações de apenas alguns segundos.

Com os dados básicos no formato inicial necessário foi possível extrair informações úteis sobre o consumo de cada circuito da casa.

ANÁLISE DESCRITIVA DOS DADOS

A análise descritiva propõe a compreensão dos dados, do passado, ou em tempo real do que acontece. É uma maneira de visualizar os dados que estão sendo tratados, entender como se organizam, bem como extrair informações qualitativas (Datastorm, 2016).

A análise descritiva dos dados de consumo dos circuitos da casa inteligente envolveu simples relatórios (gráficos) e resultados estatísticos das informações.

Estes resultados descritivos surgem como consequência da implantação do smart grid em casas inteligentes. A implantação de smart meters (medidores inteligentes) e sensores permite o acompanhamento do perfil de consumo da casa suprindo os consumidores com muitas informações. É possível sintetizar os ganhos com isso através das respostas de algumas perguntas, o que não era possível em uma rede elétrica convencional.

- Qual o consumo médio dos circuitos (locais, cômodos, equipamentos) da minha casa?

- Como o consumo de energia varia durante um ou vários dias?

- O consumo segue um padrão, de períodos do dia, ou dias da semana?

- Qual a distribuição de consumo, onde consumo mais, quando consumo mais?

- Há influência de fatores externos no consumo?

Essas análises puderam ser obtidas através de cálculos estatísticos que estão incluídos no R e são de fácil execução.

Informações de consumo médio podem ser obtidas através da função mean (média) para qualquer coluna numérica desejada. Como buscou-se informações mais elaboradas, a média foi por circuito, neste caso pelo nome do circuito, para isso foi utilizada a função aggregate, mostrada no script 7, obtendo a resposta seguinte da figura 7.

Script 7 – Utilização da função aggregate para obtenção da média de consumo por circuito.

```
aggregate(basecircuitos$RealPowerWatts~baseci
rcuitos$CircuitName,FUN = mean,baseciruitos)
```

```
cuitos$CircuitName basecircuitos$RealPowerWatts
     BedroomLights                    66.832888
    BedroomOutlets                   168.877355
      CellarLights                    30.942262
      CellarOutlets                  159.047593
   CounterOutlets1                  1025.326536
   CounterOutlets2                   513.144110
 DiningRoomOutlets                   859.440557
DisposalDishwasher                   833.514557
             Dryer                  2603.220415
       DuctHeaterHRV                    0.026400
        FridgeRange                   173.459975
```

Figura 7 – Utilização da função aggregate para obter média por circuito.

O que a função faz é associar duas colunas do dataframe obtendo assim o valor médio de RealPowerWatts agrupado por CircuitName.

As obtenções de outros resultados podem ser realizadas da mesma forma aplicando a função aggregate ou utilizando as próprias funções estatísticas para obter resultados das colunas desejadas.

Algumas destas funções são apresentadas na Tabela 1:

Tabela 1

Funções do sistema R

Função	Descrição
min	Valor Mínimo
max	Valor Máximo
median	Mediana
Função	**Descrição**
var	Variância
sd	Desvio Padrão

Funções como tapply também podem aplicar sobre algum campo um determinado fator. Por exemplo, realizamos o sumário, conforme descrito no script 8, que nos deu o resumo de várias funções estatísticas, por nome de circuitos, e podem ser verificadas na figura 8:

Script 8 – Utilização da função tapply para obter resumo de funções estatísticas por circuito.

```
tapply(basecircuitos$RealPowerWatts,basecircuitos$CircuitName,summary)
```

```
$BedroomLights
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.892   6.706   7.976  66.830 104.100 384.200

$BedroomOutlets
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.85   61.72   70.14  168.90  349.30  625.30

$CellarLights
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.295   3.041  39.100  30.940  47.500 101.000

$CellarOutlets
```

Figura 8 – Aplicação de tapply.

A função tapply, assim como a aggregate, permite a aplicação de funções estatísticas cruzando os dados de um dataframe.

Uma evolução bem-vinda da análise descritiva é a representação gráfica dos dados desejados. A representação gráfica sempre dá ao usuário uma visão melhor dos valores, além de deixar mais claras variações que não são totalmente perceptíveis em dados numéricos.

O R possuí várias funções para representar graficamente os dados de um dataframe ou resultado de operações com este. Uma informação necessária, no dataframe em análise, pode ser o consumo de determinado circuito em um período.

Para realizar de forma mais simplificada a plotagem necessária foram construídos subsets dos dados a serem analisados graficamente.

Um subset no R é uma parcela dos dados de um dataframe original baseado em algum critério de filtragem. Pode ser necessário trabalhar somente com alguns circuitos do dataframe e não todos, para isso a criação de um subset é uma forma de reduzir os dados para um novo dataframe.

Por exemplo, pôde-se criar um sub-dataframe subdados com a seguinte função do script 9:

Script 9 – Realiza a criação de um subset com dados de um dataframe.

```
subdados<-subset(basecircuitos,basecircuitos$CircuitName=="WashingMachine"||basecircuitos$CircuitName=="BedroomLights")
```

Desta forma um novo dataframe foi criado limitando-se os dados do dataframe original, para que ele possuísse somente os dados de potência do circuito que coincide com CircuitNameBedroomLights ou WashingMachine. O novo dataframe foi preenchido com os mesmos dados e tipos do original, mas foi filtrado, como indicado na função pelo CircuitName.

Agora foi possível a construção, de forma rápida, de um gráfico comparativo destes dois circuitos. No exemplo seguinte foi exibido o gráfico de linha dos dois circuitos indicados. Para isso a função ggplot foi empregada.

A função ggplot faz parte do pacote ggplot2, suas especificações podem ser encontradas na ajuda do R. Se não possuir o pacote, este pode ser instalado diretamente no R através do comando install.package('ggplot2'). Qualquer pacote existente para o R pode ser instalado desta forma.

A função subset carregou um novo dataframe com os dados dos circuitos filtrados na função, BedroomLights (Luzes do Quarto) e WashingMachine (Máquina de Lavar). Portanto para gerar o gráfico com a função ggplot, primeiramente foi carregado o pacote ggplot2 através do comando library(ggplot2), e depois executado o script 10 que criou o gráfico:

Script 10 – Gera um gráfico utilizando a função ggplot.

```
ggplot(data=subdados,aes(x=subdados$Timestamp
UTC,y=subdados$RealPowerWatts,colour=subdado
s$CircuitName)) + geom_line()
```

O gráfico apresentado na Figura 9 será gerado:

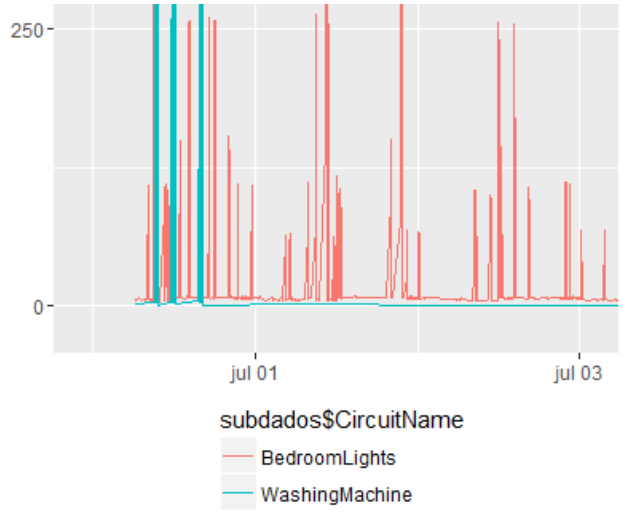

Figura 9 – Gráfico criado com o ggplot.

O gráfico foi criado de acordo a função, o eixo x com os dados de tempo, o eixo y com os dados de potência do circuito. A intenção no gráfico foi permitir uma visão do consumo dos dois circuitos para o usuário. É claro na comparação que os dados da máquina de lavar ocorrem somente em determinados dias, enquanto a potência das luzes do quarto tem elevações quase que diariamente. A análise descritiva de dados tem essa função. A análise de dados é aí mais útil para o consumidor que conseguirá enxergar se esse comportamento, dos dois circuitos, está de acordo com o real e esperado.

Este cruzamento de dados poderá fornecer informações mais interessantes, como por exemplo o consumo do ar condicionado de acordo com a variação de temperatura do ambiente, externo e interno. A análise de dados descritiva pode nos dar visões variadas do perfil de consumo elétrico de uma casa.

O R oferece ainda inúmeras formas gráficas de representação de dados. Os gráficos devem ser escolhidos de acordo com a necessidade da visualização. Além dos gráficos do pacote padrão do R, é possível através da instalação de outros pacotes gráficos utilizar versões mais aprimoradas de visualização.

Alguns exemplos de visualização gráfica no R podem ser vistos a seguir nas figuras 10 e 11:

Visualização da potência média dos circuitos conforme script 11:

Script 11 – Calcula a média dos circuitos e gera um gráfico de barras com o resultado.

```
submedia<-
aggregate(basecircuitos$RealPowerWatts~basecirc
uitos$CircuitName,FUN = mean,basecircuitos)

barplot(submedia$`basecircuitos$RealPowerWatts`,
names.arg =
submedia$`basecircuitos$CircuitName`)
```

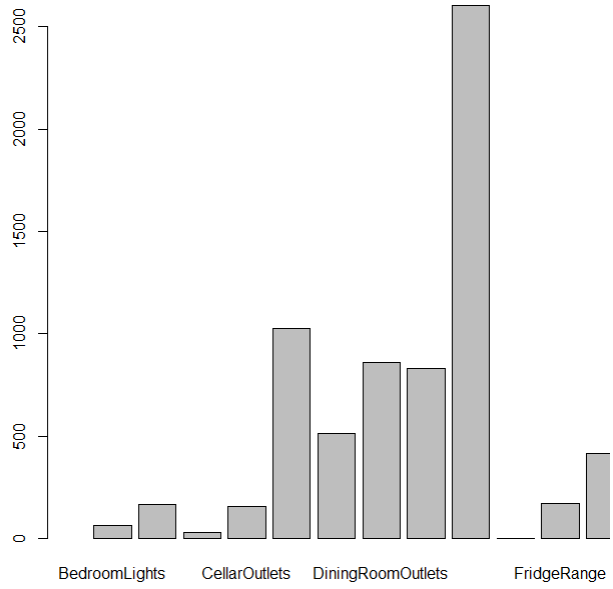

Figura 10 – Visualização do consumo médio dos circuitos.

Neste exemplo, com resultado visto na Figura 9, verifica-se de uma só vez a potência média no período de cada circuito da casa.

Potencia Aparente x Real executando o script 12 gerando o resultado na figura 10:

Script 12 – Gera gráfico cruzando os dados de consumo real e aparente para dois circuitos.

```
xyplot(RealPowerWatts ~ ApparentPowerVAs | CircuitName, groups=CircuitName, type="p", pch=16, auto.key=list(border=TRUE),

par.settings=simpleTheme(pch=16), scales=list(x=list(relation='same'), y=list(relation='same')), data=subdados)
```

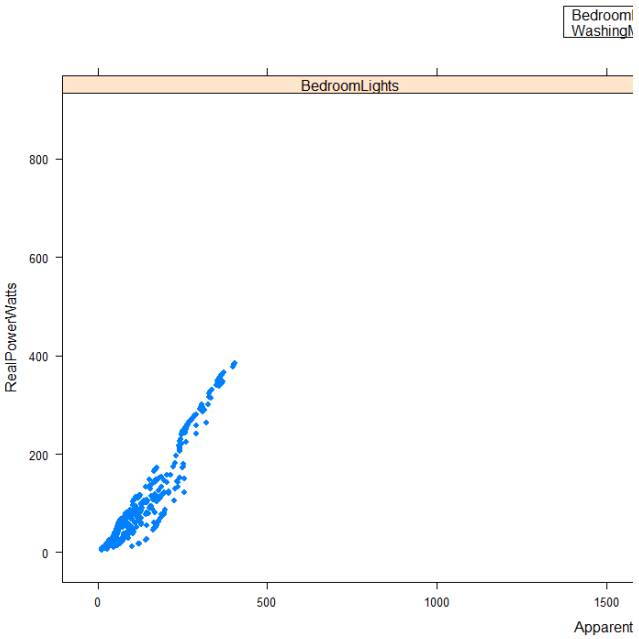

Figura 11 – Potência Aparente x Real

Neste exemplo, exibido na Figura 10, utilizando o subset subdados criado foi possível fazer o cruzamento da informação de potência aparente versus potência real do circuito, ou seja, o consumo nominal versus seu consumo real. Valores na diagonal seriam um resultado melhor, pois isso indicaria que o consumo real está próximo do nominal para aquele circuito.

A análise descritiva se dá, portanto, da inspeção dos dados passados, mostrando-se informações úteis para o consumidor. Vários usos úteis dessa análise podem ser empregados em uma casa, desde o perfil de consumo de cada parte da casa, cruzando com dados de geração de energia dado que estas casas podem ter geração própria até confrontar com os dados de consumo geral repassados pelo distribuidor.

Em uma visão mais apurada dos dados, pode-se verificar os picos de consumo de energia da casa identificando quais equipamentos e partes da casa são responsáveis por isso. De posse desses dados pode-se verificar a possibilidade de redução de consumo, ou rodizio de utilização, procurando reduzir consumo em horários de pico, onde a tarifa é maior. Neste caso a

distribuidora precisa trabalhar com bandeiras tarifárias onde isso seja contemplado.

O R se adequa bem a análise de dados descritiva pois além de possuir funções próprias para esse trabalho pode ser abastecido com pacotes que suprirão ainda mais as necessidades, além de poder ser programável para atender análise de dados ainda não disponíveis.

ANÁLISE PREDITIVA DE DADOS

A análise preditiva tem o intuito de identificar padrões nos dados do passado e realizar uma previsão do comportamento futuro daquela variável em um dado horizonte.

Com a análise dos dados passados, medidos nos circuitos da rede da casa inteligente, e o emprego de modelos estatísticos e de previsão, pode-se empregar técnicas para prever o consumo da casa, de partes dela, bem como a previsão de outros fatores que estejam disponíveis como dados do ambiente e geração de energia alternativa, em casas que possuem geração solar ou eólica. A intenção é permitir ao consumidor gerenciar melhor como será seu consumo futuro, planejar economia por decisão pessoal ou impostas pelo sistema de fornecimento e distribuição.

Na realização destas previsões foram utilizados os dados do passado, potência dos circuitos e, modelos que necessitam que os dados estejam em séries temporais.

Foram realizadas previsões de formas mais simples com a utilização de média móvel e amortecimento exponencial simples. O modelo ARIMA que é mais bem elaborado também foi utilizado, mas de forma simples sem considerar todas suas possiblidades.

As séries temporais são sequencias de observações sobre uma variável, observações em tempos discretos, usualmente equidistantes[10]. Os dados de medição de potência dos circuitos foram coletados em pequenos intervalos de segundos, mas nem sempre esses dados estavam em períodos constantes. É mais interessante podermos realizar a previsão para o horizonte de alguns dias ou meses, para isso foram necessárias algumas reduções e transformações nos dados e para prever o consumo de energia elétrica da casa.

Para facilitar a análise foi criado um novo dataframe com os dados do circuito para realizar a previsão.

No dataframe base circuitos foi incluída uma nova coluna, com o script 13, como exibido na Figura 12, para guardar a data da medição em formato sem as horas, pois foram necessários os dados médios de potência diária.

Script 13 –Cria uma nova coluna no data basecircuitos com a data em novo formato.

```
basecircuitos$datadia<-
as.Date(basecircuitos$TimestampUTC)
```

alPowerWatts	ApparentPowerVAs	datadia
1004.972	1053.855	2012-06-30
1.880	2.854	2012-06-30
10.582	15.326	2012-06-30
70.698	92.852	2012-06-30
0.594	1.553	2012-06-30

Figura 12 – Criação da coluna datadia.

A função do script acima criou a nova coluna, guardando somente a data. Após isso foi criado um subset com a média de consumo dos dados, com o script 14, seguindo os dois passos seguintes:

Script 14 – Cria um novo dataframe com a média da potência real dos circuitos do dataframe basecircuitos.

```
xgrid<-
subset(basecircuitos,basecircuitos$CircuitName=="
Grid")

xgrid<-
aggregate(xgrid$RealPowerWatts~xgrid$datadia,
FUN=mean ,data=xgrid)
```

O resultado destas duas funções foi um novo dataframe com os dados do circuito Grid, logo depois o mesmo dataframe foi substituído pelos dados do circuito Grid em média por dia. O resultado esperado pode ser observado na Figura 13.

Filter

	xgrid$datadia	xgrid$RealPowerWatts
1	2012-06-30	3158.0109
2	2012-07-01	1894.6735
3	2012-07-02	2134.0443
4	2012-07-03	1998.0896
5	2012-07-04	2380.4519
6	2012-07-05	1691.6003
7	2012-07-06	1531.8740

Figura 13 – Criação do subset com média de produção diária.

De posse desses dados pôde-se começar a trabalhar com os modelos de previsão. O subset criado envolve 11 dias de potência do circuito Grid, pode ser um período pequeno para conseguir uma previsão aderente, mas para fins de demonstração da ferramenta atendeu perfeitamente.

Uma forma de se obter a previsão é aplicar sobre esses dados modelos automáticos com ajuda de simples programas de computador, ou modelos mais elaborados que exigirão conhecimento especializado para configuração e consequente elaboração da previsão.

Para séries bem modeladas, ou seja, que variam pouco, alguns métodos mais simples podem ser empregados.

Entre os métodos mais simples existem o método ingênuo, onde se assume que o último valor de uma variável será a próxima previsão, método de média móvel e o método de amortecimento exponencial que serão apresentados a seguir.

PREVISÃO COM MÉDIA MÓVEL

Consiste em realizar uma média aritmética de um determinado período. Por exemplo, para os dados de potência analisado pode-se considerar que a previsão provável para algum dia será a média dos três últimos dias anteriores desta variável, ou seja a potência média do circuito. De forma geral a média móvel obedece a seguinte Equação 1:

$$M_t = \frac{Z_{t-1}+\cdots+Z_{t-o}}{o} (1)$$

Onde M_t é a média da variável Z dos últimos dias considerados para observação. Ou seja, podemos considerar a previsão do dia Mt, como a média dos últimos *o* dias anteriores, claro que deveremos ter os dias anteriores realizados.

Pôde-se facilmente criar uma média móvel para mostrar qual o comportamento desta previsão comparado ao real, conforme visto no script 15.

Script 15 – Realiza o cálculo da média móvel e plotagem do gráfico do resultado.

```
real<-ts(xgrid$`xgrid$RealPowerWatts`,frequency = 1)

previsão<-ma(real,order=2)

ts.plot(real,previsão,col=1:2,main="Potência real x média móvel de 2 dias")
```

No script acima foi realizada a transformação em série temporal do dado Potência de xgrid, logo depois foi criado um dado de previsão com a média móvel desses dados, com uma ordem 2, ou seja, média de 2 dias para trás. Logo depois foi utilizada a função ts.plot para visualizar graficamente esses dados gerados. O resultado da Figura 14 é gerado:

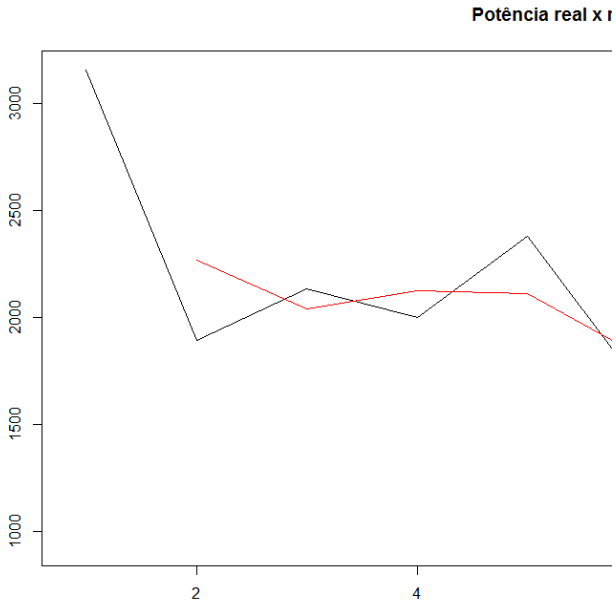

Figura 14 – Previsão com média móvel.

O gráfico mostra que para variáveis mais constantes pode-se adotar a média móvel para uma previsão em um pequeno horizonte. Em preto é exibido os dados reais de potência do circuito e em vermelho a média móvel em comparação.

AMORTECIMENTO EXPONENCIAL SIMPLES

O amortecimento exponencial pondera as observações passadas com pesos decrescentes exponencialmente para previsão de valores futuros, conforme mostrado na Equação 2:

$$Z_t = \alpha Z_{t-1} + (1-\alpha) Z_{t-1} (2)$$

Sendo 0<α<1, α é a constante de amortecimento[10].

No R é possível realizar previsões com amortecimento exponencial simples facilmente com alguns passos. Para o exemplo foram inferidos dois valores para alpha para efeitos de comparação, criando desta forma duas hipóteses. As funções para cálculo do amortecimento e o resultado são apresentados abaixo no script 16:

Script 16 – Realiza o cálculo de amortecimento exponencial e exibe gráfico com previsão.

```
hip1<-ses(real, alpha=0.2, initial="simple",h=3)

hip2<-ses(real, alpha=0.5, initial="simple",h=3)

plot(hip1,plot.conf=FALSE, ylab="Potência", main="Amortecimento Exponencial", fcol="white", type="o")

lines(fitted(hip1),col="blue",type="o")

lines(fitted(hip2),col="red",type="o")

lines(hip1$mean,col="blue",type="o")

lines(hip2$mean,col="red",type="o")
```

Os dados hip1 e hip2 recebem o amortecimento exponencial, a função que calcula o amortecimento é a ses(), logo após os dados originais são plotados, seguidos das inserções das linhas das hip1 e hip2 e suas respectivas médias previstas para 3 dias. A Figura 15 apresenta o resultado desta previsão.

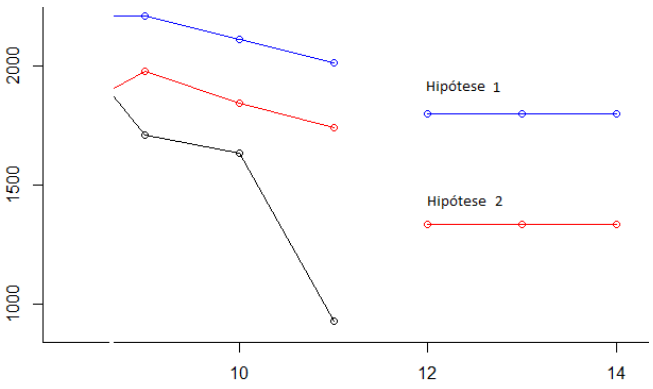

Figura 15 – Previsão com amortecimento exponencial.

Estes métodos simples de previsão podem ser aplicado para variáveis com variações mais amenas. Para projeções mais elaboradas modelos mais complexos podem ser aplicados no próprio R.

O MODELO ARIMA

Os modelos auto regressivos integrados de médias móveis (ARIMA em inglês) geram previsões através da informação contida na própria série cronológica

(HANKE; WICHERN, 2008). Os fundamentos teóricos destes modelos são bastante sofisticados, mas é possível um entendimento da essência da metodologia.

Existe uma grande variedade de modelos ARIMA. Em geral sem a componente de sazonalidade é conhecido com ARIMA (p,d,q) onde AR p é a ordem da componente auto regressiva, I d é o grau de diferenciação e MA q é a ordem da componente média móvel (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1998).

No R, para simplificação, pode-se realizar alguns comandos para se ter um modelo ARIMA auto estimado, conforme script 17, a fim de obter uma previsão para alguns dados temporais.

Script 17 – Realiza estimativa com o modelo autoarima e plota o resultado.

```
arima.estimado<-auto.arima(real)

previsoes<-forecast(arima.estimado,h=5)

plot(previsões)
```

A Figura. 16 mostra o resultado da estimativa de previsão para 5 dias com o auto arima. Os dados da figura foram reduzidos para permitir a exibição.

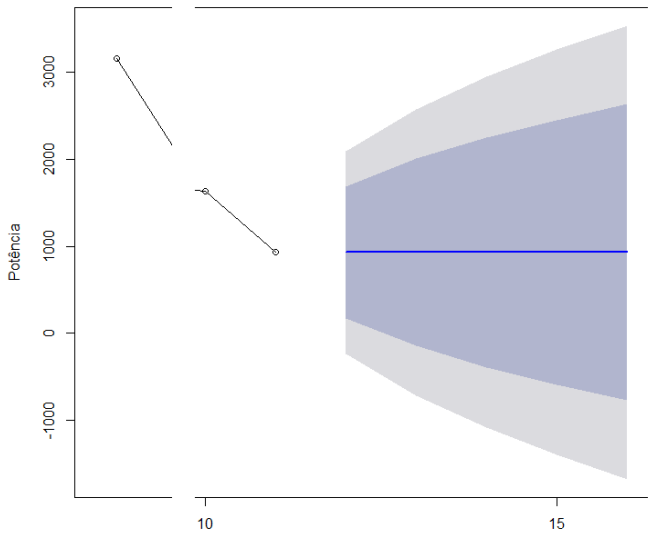

Figura 16 – Previsão com ARIMA

CONCLUSÃO

Uma casa inteligente apresenta na implantação vários desafios em diversas áreas de tecnologia. A análise de grandes volumes de dados é uma delas. Diante dessa nova visão da rede elétrica é possível, com ferramentas e conceitos já existentes, obter resultados esperados de análise e tratamento de dados como mostrado.

A utilização da linguagem R mostrou-se capaz de atender necessidades básicas tanto de mineração de dados, quanto de análises descritivas e preditivas. Principalmente na análise mostrou-se com grande capacidade de trabalhar com muitas informações, possuir suporte de vários pacotes para análises estatísticas e não ter limitações por se tratar de uma linguagem estatística que pode gerar programas para se adaptarem a novas análises.

As análises realizadas permitiram a verificação do padrão de consumo de dos equipamentos de uma residência. A análise descritiva mostrou numericamente e graficamente as variações no consumo de circuitos da casa. Já na predição foi possível utilizar modelos sobre os padrões existentes para criar curvas de previsão de consumo.

As análises foram realizadas para alguns equipamentos da casa inteligente, mas podem ser ampliadas para abranger todo o consumo existente da residência, ou até mesmo para uso em outras áreas que possuam monitoramento do consumo de energia.

Os resultados mostram que ferramentas open source já existentes podem atender o que hoje é necessário para análise de grandes volumes de dados de uma casa inteligente, e está evoluindo à medida que este novo conceito de redes inteligentes transforma a geração, transmissão e distribuição de energia do planeta.

REFERÊNCIAS

MACK, Phillippe, **Big Data, Data Mining, and Predictive Analytics and High Performance Computing**, in: JONES, Lawrence E. *Renewable Energy Integration:* Practical Management of Variability, Uncertainty and Flexibility in Power Grids, San Diego: Elsevier, 2014, Capitulo 35, p 439 – 454, ISBN: 978-0-12-407910-6.

JOHN. Hanke; WICHERN, Dean. ***Business Forecasting***. 9ª Edição. Prentice. 2008. ISBN 9780132301206.

Big Data Business. ***Conheça os 4 Tipos de Análise de Big Data Analytics*.** 2016. Disponível em: <http://www.bigdatabusiness.com.br/conheca-os-4-tipos-de-analises-de-big-data-analytics/>. Acessado em 20 abr. 2016.

Comissão Europeia. ***European Smart Grids Technology Platform*: Vision and Strategy for Europe's Electricity of the Future**. 2006. Disponível em: <https://ec.europa.eu>. Acessado em 20 abr. 2016.

HOUSSAIN, Rahat; AMANULLAH, Maung Than; ALI, Shawkat. **Evolution of Smart Grid and Some Pertinent Issues**. In: AUSTRALIAN UNIVERSITIES POWER ENGINEERING CONFERENCE (AUPEC), 20, 2010, Christchurch, New Zealand, Disponível em: <http://www.ieeeexplore.com/stamp/stamp.jsp?tp=&arnumber=5710797&isnumber=5710678>. Acessado em 20 abr. 2016.

MAKRIDAKIS, Spyros; WHEELWRIGHT, Steven; HYNDMAN, Rob. ***Forecasting*: Methods and Applications.** 3ª Edição. 1997. 656p. ISBN: 978-0-471-53233-0.

FERREIRA, Tiago Alessandro Espinola. ***Mineração e Previsão de Séries Temporais***. 2001. Recife. Disponível em: <http://www.cin.ufpe.br/~compint/aulas-IAS/kdd-012/TimeSeries.ppt>. Acessado em 20 abr. 2016.

CEMIG. ***O Que São as Redes Inteligentes de Energia?*** Disponível em: <http://www.cemig.com.br/pt-br/A_Cemig_e_o_Futuro/sustentabilidade/nossos_programas/Redes_Inteligentes/Paginas/as_redes_inteligentes.aspx>. Acessado em 03 jun. 2016.

RStudio. ***Open Source and Enterprise-Ready Professional Software for R***. Disponível em: <https://www.rstudio.com/>. Acessado em 20 abr. 2016.

U.S. Department of Energy. ***Smart Grid System Report.*** 2009. 884p. Disponível em: <http://energy.gov/oe/downloads/2009-smart-grid-system-report-july-2009>. Acessado em 20 abr. 2016

IOT AGENDA. *Smart Home or Building*. Disponível em: <http://internetofthingsagenda.techtarget.com/definition/smart-home-or-building> . Acessado em 09 jul. 2016.

EKANAYAKE, Janaka et al. ***Smart Grid*: Technology and Applications**. 1ª Edição. West Sussex: Wiley, 2012. 277p. ISBN 978-0-470-97409-4.

Datastorm. ***Tipos de Análises de Dados em Big Data*.** Disponível em: <http://datastorm.com.br/tipos-de-analise-de-dados-big-data/>. Acessado em 20 abr. 2016.

R Project. ***The R Project for Statistical Computing***. Disponível em: <https://www.r-project.org/>. Acessado em 20 abr. 2016.

LASS – Laboratory for Advanced Software Systems. ***The UMass Trace Repository. National Science Foundation***. 2009. Disponível em: <http://traces.cs.umass.edu/>. Acessado em 20 abr. 2016.

AUNG, Zeyar et al, ***Towards accurate electricity load forecasting in smart grids***, in: 4TH INTERNATIONAL CONFERENCE ON ADVANCES IN DATABASES, KNOWLEDGE, AND DATA APPLICATIONS (DBKDA), 4, 2012, Saint Gilles, *DBKDA 2012:* The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications, Saint Gilles, 2012, p 51 – 57, Disponível em: <https://www.thinkmind.org/index.php?view=instance&instance=DBKDA+2012>. Acessado em 9 jul. 2016.

FRYE, Wes. **Transforming the Electricity System to Meet Future Demand and Reduce Greenhouse Gas** Emissions. Disponível em: <http://www.cisco.com/c/dam/en_us/about/ac79/docs/Smart_Grid_WP_1124aFINAL.pdf>. Acessado em: 20 abr. 2016.