Análise e Previsão de Séries Temporais

Henrique S. Hippert

1. Introdução

1.1. Séries temporais e Estatística

As *séries temporais* são, em essência, séries de números que mostram como os valores de uma variável mudam ao longo do tempo. Gráficos destas séries atualmente estão em toda parte; por exemplo, nas páginas de um jornal, que mostram as previsões das temperaturas para cada hora do dia seguinte, ou as variações da taxa de câmbio do dólar a cada dia do último semestre.

Para os estatísticos, séries de temperaturas ou de taxas de câmbio são *amostras* dos resultados produzidos pelo complicado e sensível mecanismo que regula o clima de uma região, e pelo ainda mais complicado e sensível mecanismo que regula as finanças internacionais. Amostras não contam a história toda sobre o funcionamento destes mecanismos; apenas mostram o que aconteceu num trecho do passado. A *análise de séries temporais* é uma das áreas de Estatística, e o problema que enfrenta é o mesmo das outras áreas: examinar valores de uma variável encontrados em amostras, usando técnicas baseadas na teoria de Probabilidades, e tentar identificar os padrões de comportamento desta variável, separando aquilo que é repetitivo e previsível (o *padrão*), daquilo que é acidental e imprevisível (o *ruído, erro* ou *resíduo*). Em outros termos, tenta entender como funciona o mecanismo que gerou a amostra, e quanto num valor observado é devido ao efeito do mecanismo, e quanto é devido a influências externas desconhecidas. A *previsão*, em seguida, nada mais é que a extrapolação para o futuro do padrão extraído a partir das informações passadas.

A principal diferença entre a *Análise e Previsão de Séries Temporais* (APST) e as outras áreas da Estatística está no tipo de amostras empregadas: na APST a amostra é uma série *temporal*, isto é, uma sequência de observações *ordenadas no tempo*; na Estatística, as amostras são *aleatórias* (pelo menos em princípio), isto é, um conjunto de números não ordenados entre si. Esta ordenação obrigatória dos valores separa as duas áreas de estudo: na Estatística, ela não têm importância; na APST, ela é fundamental, e não pode ser ignorada. Numa amostra de pesos de crianças nascidas em uma maternidade durante um mês, por exemplo, tanto faz mostrar os resultados ordenados pelos pesos - dos menores para os maiores, ou dos maiores para os menores -, ou pela data do nascimento da criança, ou por ordem alfabética do nome da mãe, ou sem nenhuma ordenação. Numa série de medidas horárias de temperatura, por outro lado, os resultados têm que ser mostrados sempre da mesma forma, estritamente cronológica.

Esta diferença no tipo de amostras usadas faz com que as técnicas e ferramentas criadas pela Estatística nem sempre possam ser usadas em sua forma original. Medidas como a média e a variância, modelos como a curva de Gauss, e testes como o *t*, podem ser usados em APST. Algumas ferramentas contudo têm que ser adaptadas, como os modelos de regressão para séries temporais (por exemplo, a *regressão dinâmica*) e outras têm que ser criadas especificamente para STs (por exemplo, a medida de *autocorrelação*, derivada da correlação linear de Pearson).

Além disso, há outra conseqüência importante da diferença do tipo de amostras usadas em cada caso: valores faltantes numa amostra (*missing values*) ou valores registrados erroneamente (*wrong values*) provavelmente serão mais prejudiciais em uma ST do que numa amostra aleatória. Na Estatística, valores faltantes ou errados (se não forem extremos) podem influenciar pouco na estimação dos parâmetros de um população; na APST, porém, valores faltante impossibilitam a previsão, e valores errados podem levam a previsões totalmente erradas. Daí, a grande importância da análise exploratória e do tratamento dos dados (*data cleaning*), em APST.

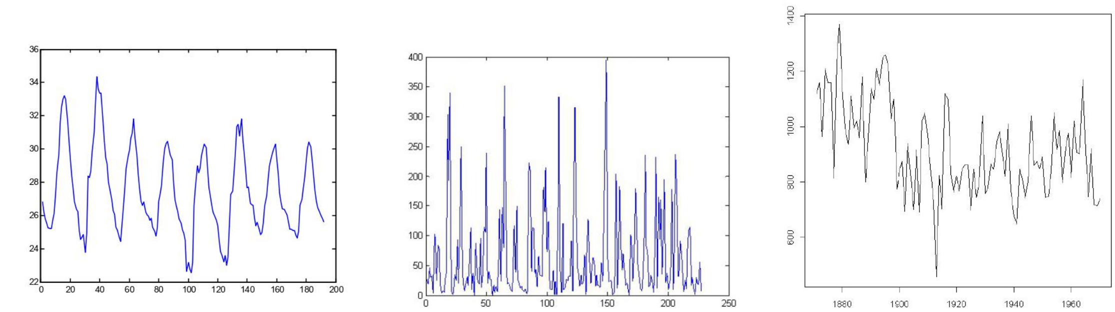
1.2. Tipos de séries temporais

As séries temporais podem ser classificadas com *discretas* ou *contínuas*. As séries discretas contém observações feitas geralmente em intervalos de tempo eqüidistantes. Algumas séries são naturalmente discretas, como por exemplo as de salários pagos a trabalhadores (uma vez por semana, ou uma vez por mês), ou de pagamentos de dividendos de um ação (geralmente trimestral ou semestral). Outras são contínuas, mas podem ser *discretizadas*, por *agregação* ou por *amostragem*. Dizemos que uma série foi discretizada por agregação quando seus valores representam a soma (ou integração) dos valores observados a cada instante ao longe de um período. Um exemplo são as séries de vendas mensais de um produto, calculadas pela soma de todas as vendas diárias feitas durante um mês. Dizemos que uma série foi discretizada por amostragem quando contém medidas pontuais de uma variável contínua feitas em intervalos iguais de tempo. Exemplos disto são as séries de temperaturas horárias em uma cidade, ou de cotações de fechamento de ações da bolsa de valores. Tanto a temperatura do ar quanto a cotação na bolsa variam continuamente ao longo do dia, mas as séries mostram apenas as medições a cada hora (temperatura), ou a última cotação de cada dia (ações). O intervalo de amostragem deve ser escolhido de acordo com o uso que será dado à série; uma série de cotações de fechamento da bolsa, por exemplo, não é útil para investidores interessados em *day trading* (estratégia de investimento baseada na compra e venda de ações num mesmo dia), mas pode ser útil para mostrar a evolução do mercado ao longo do ano. Para séries muito voláteis (isto é, que mostram grande variação), o intervalo deve ser o menor possível, para dar uma boa aproximação da variável contínua original. O processamento de sinais (por exemplo, em telecomunicações) é um problema especial de análise de ST, no qual sinais contínuos são digitalizados por conversores analógico – digital, usando intervalos de amostragem extremamente reduzidos (num CD, por exemplo, uma gravação musical é digitalizada na frequência de 44100 medições por segundo). Existem técnicas para tratar de séries contínuas (sem discretização); estas técnicas, chamadas de *análise no domínio da frequência*, consideram a série como um *sinal*, e decompõem este sinal em suas frequências componentes. Estas técnicas, que se originaram principalmente nas pesquisas de eletrônica e telecomunicações, não serão vistas neste livro.

Uma conseqüência interessante do fato de que as séries estudadas nos livros de APST são quase sempre discretas, é que a matemática envolvida usa técnicas menos familiares aos estudantes do que as da matemática contínua: em vez de derivada e integrais, por exemplo, são usados *diferenciais* e *somatórios infinitos* (veremos isto na seção 10).

1.3. De onde vêm as séries temporais

Séries temporais são estudadas em várias áreas do conhecimento. Variáveis relacionadas a fenômenos físicos, por exemplo, podem ser medidas continuamente e depois discretizadas. Exemplos comuns disto estão na meteorologia: A Fig.**1A** mostra uma série horária de temperaturas do ar no Rio de Janeiro; a Fig. **1B**, a precipitação mensal numa região da Bahia; a Fig. **1C** mostra as vazões anuais do rio Nilo.



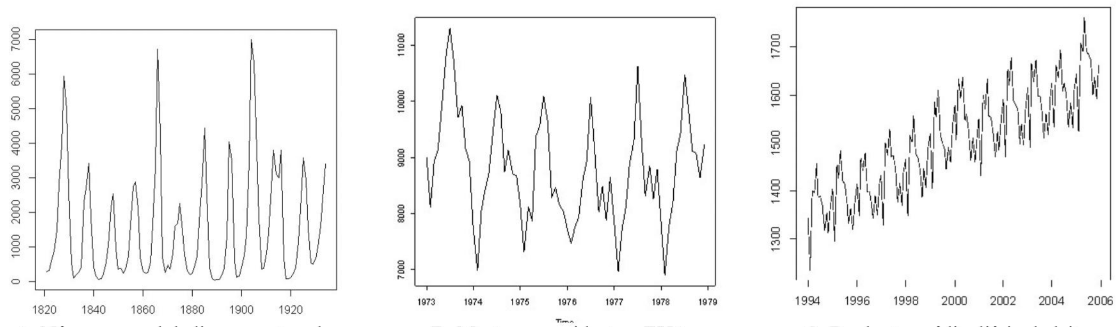
A. Temperaturas horárias, Rio de Janeiro (total de 8 dias)

B. Precipitação mensal numa região (227 observações)

C. Vazão anual do Nilo, 1871-1970 (série *Nile*)

Figura 1. Exemplos de séries naturais relacionadas a fenômenos naturais

Séries temporais também são encontradas no estudo de populações animais (a Fig. **2A** mostra o número de linces capturados no Canadá, 1821-1934), no estudo de fenômenos sociais (a Fig. **2B** mostra a número de mortes por acidentes, EUA, 1973-1978) ou de qualquer área de atividade (a Fig. **2C** mostra a produção média diária de leite por vaca, EUA).



A. Número anual de linces capturados, Canadá, 1821-1934 (série *lynx*)

B. Mortes em acidentes, EUA, 1973-1978 (série *USAccDeaths*)

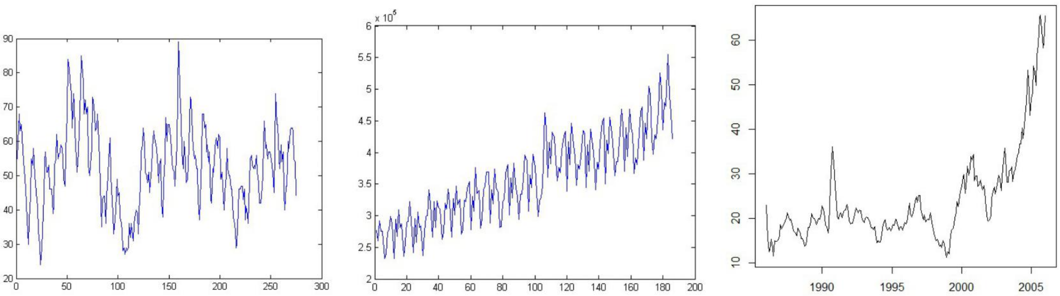
C. Produção média diária de leite por vaca, 1994-2006 (série *milk*)

Figura 2. Exemplos de séries temporais

As séries encontradas com mais frequência na mídia, contudo, são as séries econômicas e financeiras. A análise e a previsão de STs são essenciais para o planejamento de qualquer empresa – o administrador não pode planejar se não entende como funciona o mercado, e se não consegue prever com alguma precisão que demanda haverá para o produto ou serviço que oferece, e que recursos estarão disponíveis para a produção.

Séries de vendas ou de consumo geralmente variam ao longo do ano seguindo padrões bem marcados (exibem *sazonalidade*, ver seção 2.2.3), e podem ser previstas com boa aproximação, pelo menos a curto prazo. A Fig. **3A** mostra uma série de vendas mensais de casas, nos EUA, a Fig. **3B** mostra o número de quartos ocupados por noite em hotéis e similares na Austrália. Séries de preços, por outro lado, geralmente não mostram nenhum padrão repetitivo, e são extremamente difíceis de prever; um exemplo é a série de preços do petróleo, na Fig. **3C**. Alguns pesquisadores acreditam que séries de preços, e séries financeiras em geral são na verdade *passeios aleatórios*, e portanto imprevisíveis; isto será discutido na seção 10.4.5.

Existe uma área da Estatística, a *Econometria*, que desenvolve modelos multivariados especiais para as séries que interessam aos economistas, baseadas em teorias sobre o funcionamento dos sistemas econômicos. Estes modelos não serão estudados neste site.



A. Vendas mensais de casas, EUA, 1973-1995. (série *hsales*)

B. Quartos de hotéis e motéis ocupados por noite, Victoria (Austrália) [1]

C. Preço mensal do petróleo (série *oil.price*)

Figura 3. Exemplos de séries relacionadas à Economia

Não serão também discutidas neste site séries originárias da *Demografia*, área de pesquisa que faz o estudo estatístico de populações, analisando seu crescimento, a variação de estatísticas vitais, etc. Neste estudo geralmente são usadas ferramentas específicas, diferentes daquelas usuais em APST.

1.4. Por que estudar as séries temporais

Séries temporais são estudadas por pesquisadores de várias áreas diferentes, como as Engenharias (especialmente a de Produção), Administração, Economia e Finanças, e Estatística. Os objetivos do estudo são, principalmente:

1) *Previsão* – os pesquisadores estudam uma série para tentar prever seus valores futuros;

2) *Análise* – os pesquisadores estudam a série para entender suas propriedades.

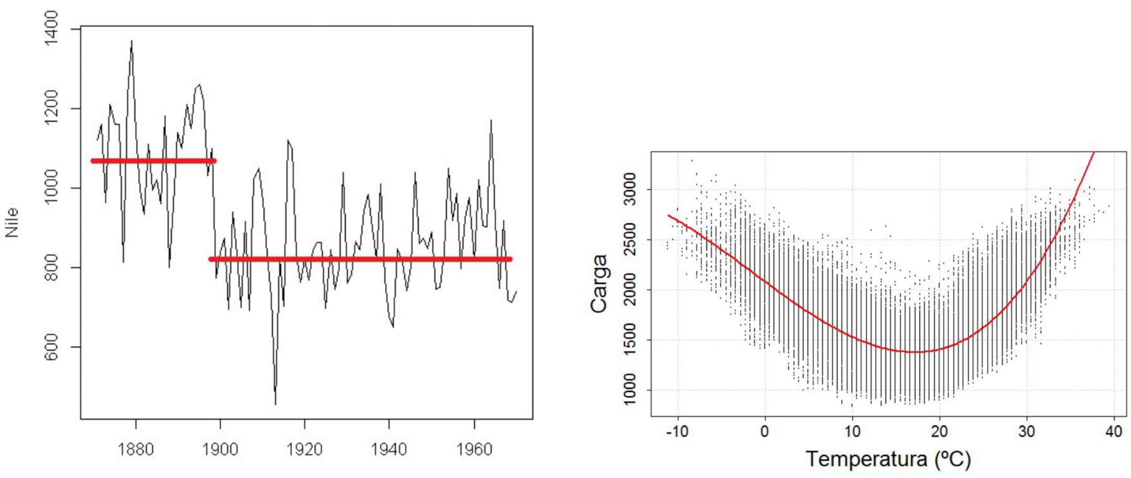
Em geral, não é possível separar totalmente estes dois motivos. Existem alguns métodos que são bastante específicos, como por exemplo, a *Decomposição Clássica*, que serve apenas para análise, mas não ajuda na previsão (seção 3). Na maioria das vezes, contudo, a análise da série pode ser uma etapa inicial de um estudo, e tem por objetivo não apenas entender as características da série, mas também sugerir um modelo que possa fazer a previsão, dependendo destas características. Mais detalhes sobre a análise e a previsão das séries temporais serão vistos nas próximas duas seções (1.4.1 e 1.4.2).

Existe ainda um terceiro objetivo de estudo de séries temporais, o *controle* de sistemas. Técnicas de controle são encontradas, por exemplo, em processos automatizados de produção, nos quais a temperatura de uma máquina é medida continuamente, e relacionada com a qualidade do produto final; modelos matemáticos são usados para controlar a máquina, a fim de ajustar a temperatura, e manter a qualidade de produto dentro de limites aceitáveis. A teoria de controle é estudada principalmente por engenheiros, usando métodos especializados, que não serão vistos neste site.

1.4.1. Análise

A *Análise* de uma série temporal tem por objetivo descrever as características da série, e decompô-la em seus componentes. Procura, por um lado, identificar os padrões de regularidade existentes, como a *tendência* de crescimento ou decrescimento dos valores ao longo do tempo, ou a *sazonalidade*, a parte da variação que ocorre regularmente a intervalos fixos (estes termos serão definidos na próximo capítulo). Por outro lado, procura também identificar e explicar as características observadas que não parecem seguir nenhum padrão regular, como os *valores discrepantes* (valores altos demais, ou baixos demais, em relação ao que era esperado) ou as *quebras estruturais*, alterações bruscas na série, causadas talvez por mudanças de legislação, crises políticas, etc.

Algumas séries tem características que podem ser facilmente discerníveis num gráfico. Por exemplo, a Fig. **4A** mostra a evolução da vazão do rio Nilo (a mesma do gráfico na Fig. **1C**). É bem evidente que houve uma quebra estrutural por volta do ano de 1900; a vazão antes oscilava em torno de z=1080, e passou a oscilar em torno de z=800; esta quebra foi causada pela construção, em 1902, da barragem de Assuã. Em outras séries, contudo, as características não são tão evidentes, e para descobri-las é preciso usar técnicas baseadas em modelos matemáticos. Veremos alguns destes modelos na seção 2.2.



A. Vazões anuais do rio Nilo (série *Nile*)

B. Carga elétrica x temperatura horárias em uma região, por um ano [2]

Figura 4. Exemplos de análise de séries temporais

A análise pode também ser feita para descrever e quantificar a variação observada na série em função da variação observada em outras séries ou variáveis externas. Por exemplo, a Fig. **4B** mostra claramente que a série de cargas elétricas horárias de uma região é dependente da série de temperaturas do ar: a carga é menor quando a temperatura está entre uns 15 ^{o}C e 20^{o}C, e tende a subir quando a temperatura está abaixo deste intervalo (devido ao uso de aquecimento elétrico) ou acima dele (devido ao ar-condicionado). Esta informação pode ser útil para a tomada de decisões administrativas, ou para embasar a escolha de um modelo de previsão para as cargas.

1.4.2. Previsão

A previsão por meio de modelos matemáticos é uma das bases da Ciência. Para testar uma teoria, cientista fazem previsões sobre os valores futuros da variável que lhes interessa, com base no modelo que desenvolveram; se estas previsões se confirmam, dizemos que a teoria recebeu *corroboração empírica*. Isto foi feito, por exemplo, por Johannes Kepler, no início do século XVII. Depois de analisar várias observações das posições dos planetas, inferiu que eles se moviam em órbitas elípticas em torno do sol. Usando os dados do passado para estimar os parâmetros destas elipses, foi capaz de prever a posição futura de um planeta de forma mais acurada do que a que era conseguida anteriormente; isto mostrou que seu modelo era melhor do que o de Copérnico, baseado em órbitas circulares.

Esta previsão (como todas as feitas na Física clássica) se baseou em modelos *determinísticos*, isto é, modelos que não incluem uma parcela de aleatoriedade em sua formulação. Exemplos destes modelos são bem conhecidos por qualquer pessoa que já estudou um pouco de Física, como por exemplo:

$$F = ma \qquad V = RI \qquad e = e_0 + v_0 t + at^2/2 \qquad E = mc^2$$

Modelos determinísticos funcionaram muito bem na Física durante séculos porque a parcela de variação aleatória não previsível nestes problemas é relativamente pequena e pode ser descartada, ou eliminada por meio de técnicas simples. As previsões feitas então por estes modelos têm resultados surpreendentemente acurados. Os astrônomos, por exemplo, prevêem a cada dia o horário do nascer e do por do sol com precisão de segundos, e eventos regulares como as fases da lua, ou as eclipses do sol com meses de antecedência. Por causa desta precisão, a Física e outras ciências naturais já foram no passado chamadas de "Ciências Exatas". Em áreas afins, como a Meteorologia, a incerteza é bem maior que na Astronomia; mesmo assim, as conhecimento têm avançado muito, e hoje é geralmente possível prever as temperaturas para cada hora do dia seguinte com erros de apenas um grau centígrado.

Em áreas como Medicina, Biologia, Ciências Sociais, Economia e Finanças, contudo, a incerteza é sempre muito grande, e os problemas têm que ser descritos em termos probabilísticos. Modelos *probabilísticos* são aqueles que incluem necessariamente um termo aleatório; por exemplo, o modelo clássico de regressão linear dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

no qual a parcela ε_i representa uma variável aleatória. Por causa desta parcela, o valor de Y_i nunca poderá ser previsto com certeza em função de X_i, mas apenas de forma aproximada.

Os modelos de previsão que iremos estudar serão sempre probabilísticos. Um exemplo simples deles é o modelo *auto-regressivo de primeira ordem*, dado por:

$$Z_t = \beta_0 + \beta_1 Z_{t-1} + \varepsilon_t$$

onde o valor da variável Z no instante t é calculado por meio de uma regressão linear sobre o valor observado no instante anterior, Z_{t-1}; esta previsão nunca será exata, porque existe sempre uma parcela aleatória ε_t, que representa a incerteza existente.

Isto nos leva a considerar: o que pode ser previsto, com uma precisão aceitável, e o que não pode? Se hoje conseguimos prever a posição dos planetas com meses de antecedência, por que não podemos prever o valor de uma ação na bolsa de valores, com alguns dias de antecedência? Se as previsões da Meteorologia melhoraram tanto nos últimos anos, será que poderemos um dia prever o futuro da economia de um país, ou mais modestamente, prever o preço do petróleo um mês à frente?

Toda previsão é baseada na extrapolação para o futuro dos padrões observados no passado da série. Isto significa pressupor que estes padrões (por exemplo, tendência ao crescimento) não irá se alterar significativamente, pelo menos não a curto prazo. É claro, não existe nenhuma maneira empírica ou estatística de validar este pressuposto; no entanto, se compreendemos razoavelmente bem o funcionamento do sistema que gera a série, e não vemos nenhuma razão para acreditar que ele será afetado por algum choque no futuro próximo, estaremos provavelmente justificados em extrapolar os padrões observados.

Isto é o mesmo que dizer que o sistema que produz os dados é "pesado", isto é, de grande inércia, e que as mudanças ocorrem de forma muito lenta. Uma analogia pode ser feita entre este sistema e um barco que observamos no mar, e cuja posição daqui a meio minuto tentamos prever. Se este barco for um grande cargueiro, a previsão pode ser feita com razoável precisão, pois um barco deste tamanho não pode frear de repente, ou mudar de direção; se o barco for um pequeno caiaque, por outro lado, a previsão pode ser impossível. O consumo de energia elétrica de uma cidade grande, por exemplo, é conseqüência de um sistema de muita inércia, formado pelos hábitos da cidade, o horário em que as fábricas e o comércio funcionam, e o horário em que as pessoas vão para o trabalho ou voltam para casa, etc. Estes hábitos não mudam subitamente, de um dia para outro. Além disso, como a cidade tem centenas de milhares de habitantes, mesmo que algumas pessoas mudem de hábitos, saiam da cidade, ou resolvam não ir trabalhar naquele dia, isto pouco irá afetar o consumo total. Por isso, uma previsão com erro de 2% a 3% é alcançável. Por outro lado, prever o consumo de um pequeno edifício residencial envolve uma incerteza maior, já que a inércia é pequena; o comportamento inesperado de um único morador pode afetar o consumo total, e isto fará que qualquer previsão esteja sujeita a grandes erros.

Existem sistemas, como a Bolsa de Valores, que praticamente não têm inércia; as cotações podem subir e descer bruscamente, em questão de minutos, devido a alterações no mercado, incidentes políticos, notícias ou rumores. Isto torna estas cotações praticamente imprevisíveis, mesmo a curto prazo.

Esta conceito de "inércia" de um sistema explica também porque previsões a longo prazo em geral não funcionam: nenhum sistema irá se manter imutável por dezenas de anos, especialmente nos dias de hoje, em que as mudanças tecnológicas ocorrem de forma muito rápida. Extrapolações ingênuas a partir de tendências do passado foram a base dos romances de ficção cientifica surgida depois da Segunda Guerra Mundial. Estes livros previam um futuro com energia barata, aviões individuais (ao invés de carros), viagens interplanetárias, etc. Nada disto aconteceu. Por outro lado, as modificações mais importantes que aconteceram nas últimas décadas, que não eram simplesmente a prolongação de tendências já existentes, não foram previstas por ninguém – nenhum dos "futurólogo" do passado previu a *internet*, computadores pessoais, celulares, ou a digitalização (que revolucionou a fotografia, a gravação de sons, a TV, etc.).

Fonte dos dados:
[1] http://www.robjhyndman.com/TSDL/
[2] https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting