Avaliação Inicial e Pré-processamento de Grandes Conjuntos de Dados: Revelação de Perspetivas nas Condições Meteorológicas e Consumo de Energia

Nuno Gomes   
MEIC- Mineração de Dados em Larga Escala  
Instituto Superior Engenharia de LisboaLisboa, Portugal  
E-mail: A18364@alunos.isel.ipl.pt  
  
Ricardo Ramos   
MEIC- Mineração de Dados em Larga Escala  
Instituto Superior Engenharia de LisboaLisboa, Portugal  
E-mail: A46638@alunos.isel.ipl.pt  
  
Rafael Carvalho   
MEIC- Mineração de Dados em Larga Escala  
Instituto Superior Engenharia de LisboaLisboa, Portugal  
E-mail: A47663@alunos.isel.ipl.pt

*Resumo - O objetivo deste trabalho é demonstrar claramente a compreensão do problema de mineração de dados, selecionando e aplicando soluções adequadas ao longo do desenvolvimento do modelo de aprendizagem. Nesta primeira fase, procura-se descrever o problema específico de mineração de dados e enquadrá-lo no contexto geral desta área. Além disto, é caracterizado detalhadamente o conjunto de dados utilizado, explicando os critérios seguidos para sua construção. São indicados os problemas encontrados durante o pré-processamento dos dados e as soluções adotadas.*

# Introdução

Perante os grandes conjuntos de dados abarcando condições atmosféricas e consumos energéticos, o objetivo é extrair informações valiosas e padrões úteis. Foi realizada uma avaliação inicial para observar os conjuntos de dados, com o intuito de refinar e gerar hipóteses para a extração de informações e padrões que possam ser conclusivos em relação a um ou mais problemas formulados. Por exemplo, determinar um classificador de zonas residenciais ou industriais com base no consumo energético e/ou gerar um modelo de regressão que permita prever o consumo energético com base nas condições meteorológicas para um determinado código postal, ou mesmo antecipar aumentos de consumo com base tanto nas condições meteorológicas como nos períodos sazonais.

Após esta fase, os dados são submetidos a um pré-processamento para criar conjuntos de dados que permitam a extração das informações desejadas e treinar modelos para obter classificadores o mais robustos possível.

Num projeto de pré-processamento e preparação de dados é imperativo cumprir algumas etapas antes de tratar de qualquer tipo de tarefa de extração de características e modulação de um problema. Deve ser feita uma limpeza dos dados, removendo eventual ruído introduzido por algumas características, e corrigir observações que possam estar incompletas, ou pouco precisas. Além disto, é essencial ser possível de compreender e caracterizar o conjunto de dados disponível, descrevendo detalhadamente a sua origem, estrutura e qualidade [1].

Identificar e corrigir os problemas dos dados é crucial para garantir a qualidade e integridade dos dados analisados para o modelo a ser treinado, perspetivando uma melhor *performance* para um determinado objetivo traçado. Posteriormente, deve ser considerado o problema da maldição da dimensionalidade, e a complexidade introduzida no treino do modelo devido a este problema, pelo que devem ser aplicadas de técnicas de redução de dimensionalidade, efetuando a seleção e redução de características. A maldição de dimensionalidade demonstra que à medida que se aumenta o conjunto de *features* de um conjunto de dados, o treino de modelos de *machine learning* requer cada vez mais um maior número de amostras, mostrando-se um obstáculo na robustez do modelo. Posto isto, a redução de dimensionalidade compromete a variância dos dados originais, porém oferecendo melhores características para o treino de modelos de aprendizagem. Durante o processo da redução de dimensionalidade removem-se informações pouco relevantes ao conjunto de dados, eliminando possível ruído na fase de treino.

Neste trabalho, o processo de análise, correção, e seleção de dados para a construção de novos conjuntos de dados relevantes ao problema a modular foi realizado utilizando a linguagem de programação R. Para tal, mostrou-se essencial desenvolver código claro e organizado que execute todas as análises propostas, garantindo a reprodutibilidade e transparência dos resultados obtidos.

Neste processo, foram criados os conjuntos de dados fundamentais para o desenvolvimento de modelos de aprendizagem. Ao compreender claramente o problema, caracterizar os dados e aplicar técnicas de pré-processamento adequadas, estamos a preparar o terreno para análises mais avançadas nas fases subsequentes.

# Caracterização do conjunto de dados disponível

A caracterização define o processo de conhecer de forma mais detalhada o conjunto fornecido, nomeadamente em termos de dimensionalidade, número de amostras, que tipo de características tem, o que cada valor representa, e identificar possíveis valores ruidosos ou em falta por forma a tomar decisões de como limpar o *dataset* para treinar um modelo com apenas informação pertinente que garanta o melhor desempenho possível [2].

Neste capítulo, o grupo descreve sucintamente o significado de cada uma das *features* dos dois conjuntos de dados. O conhecimento das variáveis com as quais tratamos permite então formular possíveis perguntas que se pretendam responder ou solucionar preparando os dados para o treino de modelos de *machine learning*.

De seguida, é feita a análise estatística procurando identificar valores ruidosos, ou em falta, e compreender o significado dos mesmos, por forma a determinar se é feita imputação dos dados, e como, ou removidas as entradas.

## Dados de consumos de energia

Os dados fornecidos são relacionados ao consumo elétrico ativo, medido em quilowatt-hora para um código postal numa determinada hora ao longo de vários dias compreendidos entre o fim de 2022 e 2023.

i) Estrutura dos Dados:

Date/Time: Representa a data e hora em que a medição foi feita. Parece estar em formato ISO 8601, incluindo o fuso horário.

Date: Data da medição no formato DD/MM/AAAA.

Hour: Hora da medição, formato HH:mm

Zip Code: Primeiros quatro dígitos do código postal do local onde a medição foi feita.

Active Energy: Variável contínua, representa a energia elétrica ativa consumida, medida em kWh.

ii) Qualidade dos Dados:

Este conjunto de dados é composto por 5 *features,* com um registo de um total de aproximadamente 3.8 milhões medições de energia ativa consumida para os vários códigos postais compreendidos ao longo do país. Verifica-se que todos as entradas possuem um valor não nulo, ou NA, para todas as *features,* pelo que não foi necessário fazer qualquer tipo de imputação. A existência de valores que possam ser considerados *outliers* foi preservada, uma vez que um pico de consumo energético pode representar alguma ocasião especial, nomeadamente um tipo de evento espontâneo e que possa ser do interesse ser identificado por um modelo de *machine learning,* sendo este um dos objetivos definidos pelo grupo na preparação deste conjunto de dados. O consumo energético, por se tratar de uma *feature* contínua, terá de ser discretizada após eventual processo de *feature selection* ou *feature reduction* por forma a evitar ruído na aprendizagem do modelo a treinar numa fase posterior. Notou-se um atributo comum ao segundo conjunto de dados fornecido relativo às condições metereológicas em Lisboa (código postal 1000), sendo este a coluna do fuso horario (*timestamp*), que pode ser útil para unir ambos e avaliar o consumo energético dadas as condições metereológicas para um conjunto de códigos postais, perspetivando um modelo geral que possa fazer uma aproximação de consumo energético dadas as condições previstas.

Tem-se como observação adicional a existência de *features* que demonstram as mesmas características, como as do fuso horário, que podem então ser tratadas apenas por uma, reduzindo *à priori* o conjunto de dados para uma dimensionalidade menor, sem que sejam necessárias técnicas de *feature selection* ou *feature reduction*.

Desta forma, apenas por compreensão e análise dos dados deste conjunto, o grupo traça três tipos de modelos que pretende treinar e avaliar a performance numa fase posterior: Dois classificadores, o primeiro responsável por identificar eventos num determinado Zip.Code com base em picos esporádicos de consumo, e um modelo classificador de Zip.Code como residencial ou industrial. O terceiro modelo trata-se de um algoritmo baseado em regressão, que com base nas condições metereológicas prevê o consumo energético para Lisboa.

## Dados de condições climatéricas

O segundo conjunto de dados retrata uma série temporal de observações meteorológicas, provenientes de uma estação meteorológica local ou de uma fonte semelhante que regista informações climatéricas em Lisboa.

i) Estrutura dos Dados:

Cada linha representa uma observação de hora a hora, com os seguintes parâmetros/colunas:

name: Nome da localidade (constantemente "Lisbon" neste conjunto).

datetime: Data e hora da observação no formato "AAAA-MM-DDTHH:MM:SS".

temp: Temperatura média em graus Celsius, variável continua.

feelslike: Sensação térmica em graus Celsius, variável continua

dew: Ponto de orvalho em graus Celsius, variável continua

humidity: Humidade relativa em percentagem representada por variável contínua.

precip: Precipitação em milímetros, variável continua

precipprob: Probabilidade de precipitação em percentagem, variável continua.

preciptype: Tipo de precipitação (por exemplo, "rain" para chuva).

snow: Quantidade de neve em milímetros, variável continua

snowdepth: Profundidade da neve em milímetros, variável continua

windgust: Rajada máxima do vento em quilómetros por hora.

windspeed: Velocidade do vento em quilómetros por hora medida a 10m do solo, variável continua

winddir: Direção do vento em graus, variável discreta

sealevelpressure: Pressão atmosférica ao nível do mar em milibares, variável continua

cloudcover: Cobertura de nuvens no céu, em percentagem, variável continua

visibility: Visibilidade em quilómetros, variável continua

solarradiation: Radiação solar em Watts por metro quadrado, variável continua

solarenergy: Energia solar em Joules por metro quadrado, variável continua

uvindex: Índice UV, variável discreta

severerisk: Risco severo, variável discreta

conditions: Condições meteorológicas resumidas.

icon: Ícone representando as condições meteorológicas.

stations: Lista de estações meteorológicas.

ii) Qualidade dos Dados:

O conjunto de dados original é fornecido pela Visual Crossing Weather [3] e apresenta uma dimensionalidade de 24 *features* com 9504 observações. Da análise estatística de observações, notamos que para a *feature ‘severerisk’* estão em falta 8760 valores, representativo de aproximadamente 92% das observações têm valores em falta. Desta forma, é afastada a ideia de remover estas observações, uma vez que a perda de informação seria consideravelmente grande, reduzindo o *dataset* a apenas 8% da dimensão original. Em simultâneo, outras *features* apresentam valores em falta, sendo elas *‘preciptype’, ‘solarradiation’* e *‘uvindex’*.

Da fonte dos dados, destacamos um parágrafo onde é relatado que observações com valores de *features* em falta não são o mesmo que o valor 0, isto é, representam valores desconhecidos ou não disponíveis na altura da medida.

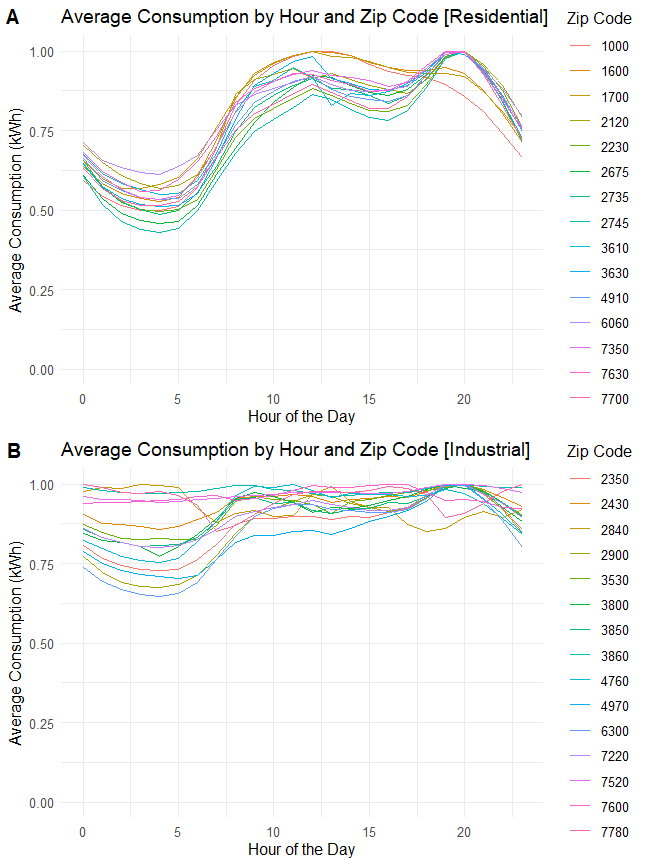
Determinada a necessidade de realizar a imputação, foi necessário definir um critério por forma a reduzir ao máximo o *bias* introduzido pela mesma. É então pertinente realizar a análise dos valores da média e desvio padrão destas *features,* onde se sabe, de boa prática, que a imputação de valores com a média da *feature* produz bons resultados e menor *bias* caso o desvio padrão seja menor, o que apenas se verificou para a *feature ‘solarradiation’.* Para a *feature ‘preciptype’*, tomou-se a decisão de imputar o valor ‘*unknown’*, indicativo que não existe informação acerca da precipitação no momento da medição da amostra.

Atendendo ao elevado desvio padrão da *feature ‘uvindex’*, a decisão de imputação passou por utilizar a mediana, o que pode resultar na introdução de *bias* para este valor. A fase de limpeza dos dados termina com a imputação do valor 0 para a *feature ‘severerrisk’* uma vez que ao observar registamos que ou tem valor 10, ou é omisso.

Desta forma, o grupo contactou com três diferentes metodologias de imputação, compreendendo os cenários onde cada uma é mais vantajosa que outras, e toma consciência da introdução do enviesamento destes valores no treino do modelo.

# Soluções adotadas e modelos a estudar

Resolvido o problema de características com valores em falta, a análise de ambos o conjunto de dados forneceu alguma informação válida e pertinente para a construção os três modelos mencionados anteriormente. Numa fase inicial, o grupo procurou analisar individualmente o conjunto de dados de consumo energético, procurando encontrar informação relevante que pudesse ser tratada e transformada para treinar um modelo com base nestes dados. Começando pelos códigos postais, com auxílio a dados externos indicadores de consumo industrial e doméstico, provenientes da PorData [4], o grupo considerou interessante a preparação de um *subset* dos dados originais, onde se categoriza cada código postal escolhido como Residencial ou Industrial, na perspetiva de treinar um classificador supervisionado, que possa prever o tipo de zona com base nos consumos energéticos ao longo do dia. Para suportar esta classificação, traçaram-se dois gráficos comparativos de consumo energético médio ao longo de 24h para zonas classificadas como residenciais e industriais, onde, depois de normalização, se verificou que em zonas industriais o consumo energético muito raramente desce acima de 70% do seu consumo máximo diário, ao contrário das zonas residenciais onde a tendência das horas “mortas” e a madrugada demonstra uma forte descida do consumo energético.

Explicar melhor, incluir eventuais gráficos

Ainda no *dataset* do consumo energético, o grupo achou relevante, embora ainda sem objetivo definido, preparar um conjunto de dados com a evolução do consumo médio energético ao longo dos diferentes dias da semana para zonas industriais e residenciais, onde se verificaram valores menos acentuados aos fins de semana.

Trabalho futuro: Picos de consumo

Numa segunda etapa, após observação dos *datasets* de consumo e de condições meteorológicas para Lisboa, o grupo definiu como objetivo a preparação dos dados com intenção de treinar um modelo de regressão com o propósito de prever consumos energéticos dadas as condições meteorológicas nesta região, uma vez que os dados disponíveis se focam para as condições meteorológicas no código postal 1000.

A interseção dos dois conjuntos de dados foi feita segundo o atributo comum, *timestamp*, para o código postal referente a Lisboa (1000) garantindo que as condições meteorológicas e o consumo ativo são correlacionáveis e precisas.

O resultado da união dos *datasets* evidencia a necessidade da eliminação de *features* repetidas e redundantes ao problema, tais como Zip.Code, name, datetime, stations, icon. É também necessário o tratamento da *feature* ‘*date’*, convertida em três novas *features* ‘*day’, ‘month’,‘year’*, tornando desnecessária a *feature ‘datetime’* razão pela qual foi anteriormente eliminada.

Mencionar dimensões finais dos *datasets preparados*

# Redução de dimensionalidade e discretização

No tópico de redução de dimensionalidade, para o primeiro conjunto preparado de consumos energéticos por zona, o grupo realizou uma transformação das variáveis categóricas: Na *class label*, as zonas industriais e residenciais foram transformadas em alternativas binárias, valor “1” ou “2”, respetivamente. Transformou-se também a característica de *timestamp* em quatro *features* cada uma representativa do dia, mês, ano, e hora registada, pelo que terminámos com um conjunto de dados de dimensionalidade n =240480, d = 8. Atendendo ao número de características, em que duas são únicas, nomeadamente o código postal e a *class label*, o grupo evidencia a baixa dimensionalidade, e deixa em aberto o estudo da *performance* de um modelo treinado com uma redução deste conjunto original, face à desvantagem da perda de informação relevante consoante o *threshold* definido. Foi aplicada a técnica de *Feature Selection* supervisionada, *Fisher’s Ratio* e também a técnica de *Feature Reduction* baseada em componentes principais, PCA, desta, após remoção das *features* de variáveis únicas (Zip code) e das *class labels*, para um *dataset* original com 6 *features* e um grau de relevância acumulada de 95%, obteve-se um conjunto reduzido a 5 componentes, o que levou o grupo a questionar se a perda de 5% de relevância por redução de 1 *feature* compensa em termos de *performance* de treino.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. https://www.scalablepath.com/data-science/data-preprocessing-phaseI.
2. https://stats.libretexts.org/Bookshelves/Applied\_Statistics/Book%3A\_Quantitative\_Research\_Methods\_for\_Political\_Science\_Public\_Policy\_and\_Public\_Administration\_(Jenkins-Smith\_et\_al.)/03%3A\_Exploring\_and\_Visualizing\_Data/3.01%3A\_Characterizing\_Data
3. https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/
4. https://www.pordata.pt/
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being publish**