Universidade de São Paulo

Instituto de Física

Estudo do Material Particulado em Nima-Ghana

Thiago Gomes Veríssimo

Orientador: Américo Kerr

Dissertação de mestrado apresentada ao Instituto de Física para a obtenção do título de Mestre em Física

---

Comissão examinadora:

Professor 1

Professor 2

São Paulo

2014

Aluno, Thiago Gomes Veríssimo

Poluição do Ar em Gana

100 páginas

Dissertação (Mestrado) - Instituto de Física da Universidade de São Paulo. Departamento de Física Aplicada.

1. Palavra-chave1
2. Palavra-chav2
3. Palavra-chave3

I. Universidade de São Paulo. Instituto de Física. Departamento de XXXXXXXX.

Comissão Julgadora:

Prof. Dr.
Nome

Prof. Dr.
Nome

Prof. Dr.
Américo Kerr

Dedicatória Dedicatória...

Resumo

Estuda-se a problemática da poluição do ar em Gana-Acra.

Palavras-chave: Gana, Acra, Poluição do ar

Abstract

This research looks for...

Keywords: Ghana, Accra, Air Pollution

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

PM_{10} Material Particulado Inalável com diâmetro menor ou igual à 10 $\mu g/m^3$

$PM_{2.5}$ Material Particulado Fino com diâmetro menor ou igual à 2,5 $\mu g/m^3$

EDXRF Fluorescência de raios X por dispersão em energia

FA Análise Fatorial

PMF Positive Matrix Factorizarion

SSA Sub-Saharan Africa

Conteúdo

Capítulo 1

Introdução

Vamos colocar alguma epigrape?

Saramago

1.1 África Subsariana

1.2 Acra

Acra é a capital de Gana e está localizada no Golfo Guiné. Ela tem uma área total de mais de 2500 km^2 com elevação que varia de 0 até 100 pés do nível do mar.

Período chuvovo: Abril-Julho e Setembro-Outubro. Ocorrêcnia do Harmathan: Novembro-Março com ventos para direção sudeste.

O grupo de Harvard (Arku et al., 2008) conduziu um levantamento nos níveis de poluição, bem como da distribuição espacial e temporal de alguns poluentes em duas regiões periféricas de Acra: (Dionisio et al., 2010)

- Jamestown/Ushertown: região entre a Costa e o centro comercial local.
- Nima: Centro comercial de Acra, cercada com Rodovia.

Nos dois bairros há poucas ruas pavimentadas, com exceção das principais avenidas.

Pesquisas recentes tem avaliado a poluição do ar em favelas (regiões periféricas) (Sclar et al., 2005) e (Riley et al., 2007).

A África Subsariana (SSA) é atualmente região no mundo que tem a maior taxa de transição da população rural - ainda predominante - para cidade (Montgomery, 2008).

Mesmo assim, as cidades da *SSA* ainda não possuem sistemas de monitoramento sistemático de poluição do ar e suas implicações na saúde (Ezzati et al., 2004). Além disso há poucas pesquisas acadêmicas dos níveis de poluição do ar nos países da *SSA*.

Diferente dos países industrializado, onde as principais fontes de poluição são os setores da industria e do transporte, os países da *SSA* tem como principal fonte poluidora a queima de biomassa, sendo comum o uso no cozimento de alimentos, tanto em regiões urbanas quanto rurais (Smith et al., 2004).

1.3 Poluição do Ar

Poluição do ar urbana é uma complexa mistura de emissões e naturais e antropogenicas.

O que é material particulado?

Tipos de classificações de Material Particulado: tamanho, por formação (primária ou secundária), composição química, remoção da atmosfera. (colocar a figura clássica do seinfeld de distribuição), forma da partículas (espérica, espiga, ...) ?

nucleação: condensação de vapores quentes (formando núcleo de condensação) ou durante transformação de gás em partícula. Remoção: aglomeração.

acumulação: formação: from nucleação através coagulação ou condensação de vapores.remoção: deposição seca ou úmida.

grossas: processos mecânicos: fragmentação, movimentação, manuseio. remoção: sedimentação.

Qual faixa é mais numerosa? usar aquele trabalho do Seinfeld da Fátima? Qual a composição química por faixa?

fina: íons $SO_4^=$ e NO_3^-, carbono elementar, carbono orgânico, compostos orgânicos elementar condensados?, metais (cádmio, níquel, vanádio, zinco, cromo, ferro, mercúrio), sulfatos, nitratos, nitrato amônia? o óxido de nitrogênio (NOx) e amônia NH3 formam nitrato de amônio e o dióxido de enxofre (SO2) amônia NH3 formam o sulfato de amônio, portanto são secundário.

Qual a diferença da composição entre ambiente rural e urbano. Artigo que a fátima deu seria interessante...

grossa: solo, fuligem, pólen. origem mineral: Si, Al, ferro, K, Ca e metais alcalinos.

nomenclatura adotada

relação com sistema respiratório humano (deposição em função do diâmetro da partícula). as partículas mais finas (PM2.5) chegam nos bronquíolos e as maiores (ainda no PM2.5) só

alcançam os alvéolos. e como o corpo as elimina? Macrófago alveolar ou o sistema linfático a expulsam? As maiores ficam no nariz (traqueobrônquica) nasofaringe . Os vírus são PM2.5 e bactérias são PM10, lega né?

ao depositar em folhas, as partículas impedem a absorção de luz.

visibilidade

formação de nuvens.

danos a saúde (saldiva e WHO) e ao ambiente.

Legislação nacional e internacional.

fontes naturais e antropogênicas. móveis e estacionárias. SPECIATE. As fontes depende da região.

1.4 Objetivos

Identificar e quantificar as fontes

Capítulo 2

Métodos matemáticos e procedimentos experimentais

Vamos colocar uma segunda epígrape?

Saramago

2.1 Modelos Receptores

Modelo receptor é uma abordagem matemática para quantificar o efeito das fontes nas amostras. Determinar as fontes a partir do receptor. (no site da cetesb tem um esquema legal.).

Análise multivariada reduz as dimensões (variáveis) de um conjunto de dados em um conjunto de dados analítico complexo que poderão ser interpretados como tipo de fontes. Outlier é um desafio para modelos multivariados porque ele são de infrequentes e de curta duração e com alta concentrações. O processo de identificação das fontes principais de poluição do ar é complexo, havendo diversos caminhos, os quais as técnicas estatísticas multivariadas são historicamente usadas.

Tentam identificar e quantificar as fontes de poluição baseados em medidas de poluentes no receptor. O número de fatores dependera do conhecimento do usuário sobre as fontes, número de amostras, resolução da amostragem, e espécies medidas. Modelos receptores: Regressão linear múltipla, redes neurais, cluster, edge detection, fator de enriquecimento, BQM, análises multivariadas. Quais são os mais usados em poluição do ar?

O Balanço químico de massa necessita dos perfis das fontes.

modelo de dispersão: determinar o poluente no receptor a partir de fontes conhecidas, exemplo: determinar a poluição em um ponto receptor usando um inventário mais condições

atmosféricas. O mais usado pelos governos é o da hipótese que a distribuição espacial dos poluentes dentro da pluma é gaussiana.

Os dois se complementam.

fundamentação do modelo receptor: conservação de massa. No fundo todos modelos resolvem a mesma equação:

Dada uma amostragem, coletamos j amostras válidas, isto é, removendo-se filtros danificados, contaminados ou filtros que passaram por problemas no amostrador. Esses filtros são analisados por diversas técnicas, de acordo com o interesse do pesquisador, tais como: análise gravimétrica, refletância, fluoresncência de Raios X, dentre outras. Essas análises resultam em i grandezas medidas.

$$x_{ij} = \sum_{p=1}^{P} g_{ip} f_{pj} + \epsilon_{ij} \tag{2.1}$$

x_{ij} = concentração na amostra i de espécie j. f_{pj} = concentração da espécie j emitida na fonte p. g_{ip} = contribuição da fonte p para amostra i.

ϵ = erro do modelo.

traçar um panomara da evolução histórica dos modelos receptores. Balanço químico de massas: o quanto dos perfis de fontes medidas explica as concentrações dos elementos. Balanço Químico de Massa é ideal que não haja muita mudança entre a emissão e o receptor. problema quando fontes importantes são omitidas do cálculo ou fonte incluida não representa fonte real.

supoções: composição das fontes constantes, sem especie reativas, todas fontes devem ser incluidas, numeros de fontes menor que o número de espécie,

falar um pouco do speciate.

O conhecimento do pesquisador é imprescindível para fazer o relacionamento entre fator e fonte(s), afim de avaliar o significado físico das fontes. Para tal, informações das possíveis fontes poluidoras próximas ao ponto receptor, dados meteorológicos do período de coleta, inventário de emissões e qualquer mais informação que o pesquisador julgar necessário devem ser usadas para fazer esse relacionamento.

boa rodada BQM, teste de performance (por amostra devem ser satisfeitos) 1) statistic-t > 2 2)R2 >=0,8 3) Porcentagem da Massa entre 80 e 100 4) qui-quadrado < 4 5) concentração modelada/concentração medida entre 0,2 e 2,0. (como calcular o Grau de Liberdade?)

problemas: alta quantidade de dados.

Redução da Dimensão Modelo receptor é uma abordagem matemática para quantificar o efeito das fontes nas amostras. Análise multivariada reduz as dimensões (variáveis) de um

conjunto de dados em um conjunto de dados analítico complexo que poderão ser interpretados como tipo de fontes.

As técnicas de redução da dimensão permitem manter a qauntidade de informação contida inicialmente nos dados. Sabemos da informação contida pela variância e covariancia (correlação) inicial dos dados. Como podemos usar a redução dos dados: 1) Combinação linear que reflete a variação dos dados. 2) Os dados iniciais estão organizados de forma revelar ium variável latente (Exploratory Factor Analysis EFA) 3) Devemos confirmar nossa crença de como os dados originais são organizados em um jeito particular (Confirmatory Factro analysis)

Fator neste caso é uma variável latente, um construct. Uma variável latente é um variável que não pode ser medida diretamente, mas sim através de outras variáveis obseváveis

O ponto inical é a matrix de correlação: fazer uma inspeção. O olhometro é um bom método para validar nossa análise. Mas temos outros: 1) Bartlet Test of sphericity: Compara nossa matriz de correlaçao com a identidade (correlação zero), fornece p-value. Se p-value for pequeno indica que é improvável obetermos essa matrix de correlação de uma população com correlação zero. Problema: o contrário não verdade. (Norman Streiner) 2) Kaiser -Meyer-Olkin measure of sampling adequacy: Não produz p-value, Remover valores menores que 0.7. Esse testes indicam que podemos validar nossa factor analises.

Outra problema da AF: Multicolinearidade (Variavel A = variavel B + 5). Pare decobrir verificar o determiante da matriz de correlação e ver se é maior que 0.00001. Ou excluir se a correlação entre duas for 1.

Se todas correlação forem abaixo 0.3, não teremos resultado.

Critério para quantificar quantas variáveis latente queremos: autovalor maior que 1 ou ...

Factor loading: Correlação entre a variável observada e o fator. é similar ao coeficiente padronizado da regressão multipla.

Comunalidade: A influencia total numa variavel de todos os fatores relacionadas com ela. é igual a soma dos factor loadings ao quadradro desse fatores, e fazendo uma analogiua com a regressão múltipla, é a variancia explicada R $\hat{2}$, podemos representar em prcentagem (% da variabilidade prevista pelo modelo) . 0 indica que a variável não foi nada explicada pelos fatores e 1 indica que foi completamente explicada pelos fatores.

singularidade: Porção da variabilidade de uma variavel que não pode ser explicada pelas variáveis latentes relacionadas a essa variavel. 1-conunalidade; Ou seja, é a porcentagem da variabilidade que não pode ser prevista pelo modelo.

Total da variância explicada: Quanto da variabilidade dos dados foi modelada pelas variáveis latentes.

rotação: maximiza os factor loading para melhor enetendiemnto dos resultados. Vaimax: afirma que os fatores não são correlacionados.Promax: os fatores podem estar correlacionados, da a correlação entre os fatores.

Montar fator: excluir factor loading menores que 0.3/0.4 ou fazer p-value;

reificação (transformação de algo abstrato em concreto): Fator é alhuma coisa na realidade? Há evidencias que a variável latente existe?

Factor Scores: calculo: variavel dependente: fator score previsto e a independente é variavel obrservada. Matrix de coeficiente dos factor score, salvar valores padronizados.

$$FatorScore1 = coeficiente_{elemento1fator1} * valor_elemento1 + coeficiente_{elemento2fator1} * valor_elemento2... \tag{2.2}$$

Fazemos isso para cada amostra e temos o factor score para cada amostra. Lembrar que no PCA não perdemos variância no processo de extração, em outros sim, só podemos fazer isso no PCA, em outros casos o programa estima. Verificar se a média do factor score é zero e o desvio padrão 1. Plotar factor score 1 pelo 2, não devem ser correlacionados. Rotacionar com promax e ver o os factor scores. Como os factor scores são normalizados o que o valor de -3 sugere?

Podemos fazer um diagrama

Everitt & Dunn 2001 page 288: Hills (1977) has gone as far as to suggest that factor analysis is not worth the time necessary to understand it and carry i t out. And Chatfield and Collins (1980) recommend that factor analysis should not be used in most practical situations. Such criticisms go too far. Factor analysis is simply an additional, and at times very useful, tool for investigating particular feature s of the structure of multivariate observations. Of course, like many models used in analysing data, the one used in factor analysis is likely to be only a very idealized approximation to the truth in the situations in which it is generally applied. Such a n approximation may, however, prove a valuable starting point for further investigations

Olhar a comunalidade da PCA e FA, PCa começa com1 , FA não

Comemoração de 100 anos de Analises de fatores http://www.fa100.info/down.htm

fatores absolutos

$$FA = \frac{L_{ij}\sigma_i}{\sigma_{PM}L_{PM_j}} \tag{2.3}$$

L = loadings i = espécies j = fatores extraídos

Capítulo 3

Análise multivariada dos dados

3.1 Visão Geral

For this tutorial, we will assume that the appropriate number of factors has already been determined to be 2, such as through eigenvalues, scree tests, and a priori considerations. Most often, you will want to test solutions above and below the determined amount to ensure the optimal number of factors was selected.

Note that several rotation and factoring methods are available when conducting EFA. Rotation methods can be described as orthogonal, which do not allow the resulting factors to be correlated, and oblique, which do allow the resulting factors to be correlated. Factoring methods can be described as common, which are used when the goal is to better describe data, and component, which are used when the goal is to reduce the amount of data. The fa() function is used for common factoring. For component analysis, see princomp(). The best methods will vary by circumstance and it is therefore recommended that you seek professional council in determining the optimal parameters for your future EFAs. In this tutorial, we will use oblique rotation (rotate = "oblimin"), which recognizes that there is likely to be some correlation between students' latent subject matter preference factors in the real world. We will use principal axis factoring (fm = "pa"), because we are most interested in identifying the underlying constructs in the data.

Sobre o pacote psych: desenvolvido em Northwestern University. Usa redução da dimensão por fatores, cluster ou PCA. Revelle, W. (2012). psych: Procedures for Personality and Psychological Research. North- western University, Evanston. R package version 1.2.8.

Solutions to this problem are examples of factor analysis (FA), principal components analysis (PCA), and cluster analysis (CA). All of these pro- cedures aim to reduce the

complexity of the observed data. In the case of FA, the goal is to identify fewer underlying constructs to explain the observed data. In the case of PCA, the goal can be mere data reduction, but the interpretation of components is frequently done in terms similar to those used when describing the latent variables estimated by FA.

An alternative to factor analysis, which is unfortunately frequently confused with factor analysis, is principal components analysis. Although the goals of PCA and FA are similar, PCA is a descriptive model of the data, while FA is a structural model. Psychologists typically use PCA in a manner similar to factor analysis and thus the principal function produces output that is perhaps more understandable than that produced by princomp in the stats package. Table 5 shows a PCA of the Thurstone 9 variable problem rotated using the Promax function. Note how the loadings from the factor model are similar but smaller than the principal component loadings. This is because the PCA model attempts to account for the entire variance of the correlation matrix, while FA accounts for just the common variance. This distinction becomes most important for small correlation matrices. Also note how the goodness of fit statistics, based upon the residual off diagonal elements, is much worse than the fa solution.

reduce the complexity of the data and attempt to identify homogeneous subgroupings

How many dimensions to use to represent a correlation matrix is an unsolved problem in psychometrics. There are many solutions to this problem, none of which is uniformly the best. Henry Kaiser once said that "a solution to the number-of factors problem in factor analysis is easy, that he used to make up one every morning before breakfast. But the problem, of course is to find the solution, or at least a solution that others will regard quite highly not as the best" Horn and Engstrom (1979).

1) Extracting factors until the chi square of the residual matrix is not significant. 2) Extracting factors until the change in chi square from factor n to factor n+1 is not significant. 3) Extracting factors until the eigen values of the real data are less than the corresponding eigen values of a random data set of the same size (parallel analysis) fa.parallel (Horn, 1965). 4) Plotting the magnitude of the successive eigen values and applying the scree test (a sudden drop in eigen values analogous to the change in slope seen when scrambling up the talus slope of a mountain and approaching the rock face (Cattell, 1966). 5) Extracting factors as long as they are interpretable. 6) Using the Very Structure Criterion (vss) (Revelle and Rocklin, 1979). 7) Using Wayne Velicer's Minimum Average Partial (MAP) criterion (Velicer, 1976). 8) Extracting principal components until the eigen value < 1.

O propósito da Análise de Componentes Principais (ACP) é encontrar a menor e melhor

representação dimensional da variãncia da um conjunto de dados

A ACP é um técnica multivariada que permite-nos encontrar os padrões sistemáticos da variação dos dados. Do ponto de vista de análise dos dados, a ACP é usada para estudar um tabela de observações e variáveis com a principal ideia de transformar as variáveis observadas em um conjunto de novas variáveis, as componentes principais, que são não correlacionadas e explicam a variação do dados.

Existe no R muitas funções de diferente pacotes que no permitem calcular a PCA.

Resultados do PCA: autovalores (variação dos dados), scores da PCA (estrutura das observações), loadings(correção entre as variáveis e as componentes)

Muitas vezes as variáveis no PCA estão em escala diferentes, devemos normalizar tomando: mean=0 e variace=1

Principal Components Analysis (PCA) and Common Factor Analysis (CFA) are distinct methods. Often, they produce similar results and PCA is used as the default extraction method in the SPSS Factor Analysis routines. This undoubtedly results in a lot of confusion about the distinction between the two.

The bottom line is, these are two different models, conceptually. In PCA, the components are actual orthogonal linear combinations that maximize the total variance. In FA, the factors are linear combinations that maximize the shared portion of the variance–underlying "latent constructs". That's why FA is often called "common factor analysis". FA uses a variety of optimization routines and the result, unlike PCA, depends on the optimization routine used and starting points for those routines. Simply there is not a single unique solution.

In R, the factanal() function provides CFA with a maximum likelihood extraction. So, you shouldn't expect it to reproduce an SPSS result which is based on a PCA extraction. It's simply not the same model or logic. I'm not sure if you would get the same result if you used SPSS's Maximum Likelihood extraction either as they may not use the same algorithm.

It is important to recognize that rotated principal components are not principal components (the axes associated with the eigen value decomposition) but are merely components. To point this out, unrotated principal components are labelled as PCi, while rotated PCs are now labeled as RCi (for rotated components) and obliquely transformed components as TCi (for transformed components)

My understanding is that the distinction between PCA and Factor analysis primarily is in whether there is an error term. Thus PCA can, and will, faithfully represent the data whereas factor analysis is less faithful to the data it is trained on but attempts to represent underlying trends or communality in the data. Under a standard approach PCA is not rotated,

but it is mathematically possible to do so, so people do it from time to time. I agree with the commenters in that the "meaning"of these methods is somewhat up for grabs and that it probably is wise to be sure the function you are using does what you intend - for example, as you note R has some functions that perform a different sort of PCA than users of SPSS are familiar with.

First, Principal Components Analysis (PCA) is a variable reduction technique which maximizes the amount of variance accounted for in the observed variables by a smaller group of variables called COMPONENTS. As an example, consider the following situation. Let's say, we have 500 questions on a survey we designed to measure stubbornness. We want to reduce the number of questions so that it does not take someone 3 hours to complete the survey. It would be appropriate to use PCA to reduce the number of questions by identifying and removing redundant questions. For instance, if question 122 and question 356 are virtually identical (i.e. they ask the exact same thing but in different ways), then one of them is not necessary. The PCA process allows us to reduce the number of questions or variables down to their PRINCIPAL COMPONENTS.

PCA is commonly, but very confusingly, called exploratory factor analysis (EFA). The use of the word factor in EFA is inappropriate and confusing because we are really interested in COMPONENTS, not factors. This issue is made more confusing by some software packages (e.g. PASW / SPSS, SAS) which list or use PCA under the heading factor analysis.

Second, Factor Analysis (FA) is typically used to confirm the latent factor structure for a group of measured variables. Latent factors are unobserved variables which typically can not be directly measured; but, they are assumed to cause the scores we observe on the measured or indicator variables. FA is a model based technique. It is concerned with modeling the relationships between measured variables, latent factors, and error.

As stated in O'Rourke, Hatcher, and Stepanski (2005): "Both (PCA e FA) are methods that can be used to identify groups of observed variables that tend to hang together empirically. Both procedures can also be performed with the SAS FACTOR procedure and they generally tend to provide similar results. Nonetheless, there are some important conceptual differences between principal component analysis and factor analysis that should be understood at the outset. Perhaps the most important deals with the assumption of an underlying causal structure. Factor analysis assumes that the covariation in the observed variables is due to the presence of one or more latent variables (factors) that exert causal influence on these observed variables"(p. 436).

Final thoughts. Both PCA and FA can be used as exploratory analysis. But; PCA is

predominantly used in an exploratory fashion and almost never used in a confirmatory fashion. FA can be used in an exploratory fashion, but most of the time it is used in a confirmatory fashion because it is concerned with modeling factor structure. The choice of which is used should be driven by the goals of the analyst. If you are interested in reducing the observed variables down to their principal components while maximizing the variance accounted for in the observed variables by the components, then you should be using PCA. If you are concerned with modeling the latent factors (and their relationships) which cause the scores on your observed variables, then you should be using FA.

Details

Useful for those cases where the correlation matrix is improper (perhaps because of SAPA techniques).

There are a number of data reduction techniques including principal components analysis (PCA) and factor analysis (EFA). Both PC and FA attempt to approximate a given correlation or covariance matrix of rank n with matrix of lower rank (p). nRn = nFk kFn' + U2 where k is much less than n. For principal components, the item uniqueness is assumed to be zero and all elements of the correlation or covariance matrix are fitted. That is, nRn = nFk kFn' The primary empirical difference between a components versus a factor model is the treatment of the variances for each item. Philosophically, components are weighted composites of observed variables while in the factor model, variables are weighted composites of the factors.

For a n x n correlation matrix, the n principal components completely reproduce the correlation matrix. However, if just the first k pri

It is important to recognize that rotated principal components are not principal components (the axes associated with the eigen value decomposition) but are merely components. To point this out, unrotated principal components are labelled as PCi, while rotated PCs are now labeled as RCi (for rotated components) and obliquely transformed components as TCi (for transformed components). (Thanks to Ulrike Gromping for this suggestion.)

Rotations and transformations are either part of psych (Promax and cluster), of base R (varimax), or of GPArotation (simplimax, quartimax, oblimin).

Some of the statistics reported are more appropriate for (maximum likelihood) factor analysis rather than principal components analysis, and are reported to allow comparisons with these other models.

Although for items, it is typical to find component scores by scoring the salient items (using, e.g., score.items) component scores are found by regression where the regression weights are $R^{(} - 1)$ lambda where lambda is the matrix of component loadings. The regression approach

is done to be parallel with the factor analysis function fa. The regression weights are found from the inverse of the correlation matrix times the component loadings. This has the result that the component scores are standard scores (mean=0, sd = 1) of the standardized input. A comparison to the scores from princomp shows this difference. princomp does not, by default, standardize the data matrix, nor are the components themselves standardized. By default, the regression weights are found from the Structure matrix, not the Pattern matrix.

Jolliffe (2002) discusses why the interpretation of rotated components is complicated. The approach used here is consistent with the factor analytic tradition. The correlations of the items with the component scores closely matches (as it should) the component loadings.

3.2 *PMF* - Positive Matrix Factorizarion

3.2.1 Introdução

O *Positive Matrix Factorizarion (PMF)* é outro método multivariado usado em modelos receptores para resolver a equação da conservação da massa (citar equação).

O *PMF* também é um tipo de Análise Fatorial (lincar para capitulo de analise fatorial) mas ao invés de fazer redução da dimensão dos dados decompondo a matriz de correlação em autovalores e autovetores, resolve a equação de conservação usando a resolução por mínimos quadrados. Em pesquisas de poluição atmosférica o método *PMF* tem ganhado espaço nos últimos anos, devido 3 motivos principais:

- Disponibilização de uma versão gratuita do software que implementa o método *PMF* pela *EPA*, o *PMF EPA 5.0* (Norris et al., 2014). Ainda existe a versão proprietária e comercial, com mais recursos.
- Incorporação de um algoritmo robusto desenvolvido por (citar paatero e tapper) que impede o aparecimento de valores negativos no perfil e contribuição de fontes.
- Ponderação pelas incertezas das concentrações, diminuído assim o peso de espécies com incertezas altas.

Em termos de reprodutibilidade da pesquisa científica o *PMF EPA 5.0* nos fornece os seguintes recursos:

- Fixação de *Random Seed*.
- Arquivo *XML* com as configurações usadas na rodada.

3.2.2 *PMF*, CMB e FA

Alguns trabalhamos mostram que PMF e CMB se correlacionam. Os perfis de fontes calculado pelo PMF são similares aos medidos de fato para USP no CMB (citar trabalho do Luis?).

Artigo que mostra um caso de não correlação entre os fatores extraídos pelo CMB e PMF (buscar artigo).

3.2.3 Desenvolvimento Teórico

Conservação da massa

Seja,

- c_{ij} matriz de concentração
- u_{ij} matriz de incertezas (experimentais e analíticas)
- i as amostras válidas
- j as espécies medidas

Pode-se escrever a equação da conservação da massa (citar equação) no contexto do *PMF* 3.1:

$$c_{ij} = \sum_{k=i}^{p} g_{ik} f_{kj} + e_{ij} \tag{3.1}$$

Onde,

- p: O número de fatores informado pelo usuário.
- g_{ik}: Contribuição dos fatores nas amostras (*Factor Score*).
- f_{kj}: Perfil da fonte ou assinatura da fonte, ou seja, a distribuição das espécies nos fatores. (*Factor Loadings*).
- e_{ij}: Matriz dos resíduos escalados pelas incertezas.

O objetivo do *PMF* é encontrar g_{ik} e f_{kj}. (k é um fator genérico), pois a matriz c_{ij} é decomposta em g_{ik} e f_{kj}. Assim, a pergunta que temos que fazer ao trabalha com o *PMF* é:

Quais g_{ik} *e* f_{kj} *melhor reproduzem* c_{ij}*?*

O usuário informa a quantidade de fatores desejados p e o *PMF* sempre encontrará uma solução para essa quantidade de fatores. Entretanto, é necessário avaliar a qualidade do ajuste,

seguindo os passos que serão apresentados logo a seguir, assim como interpretar o significado físico dos fatores.

Função Objeto

Uma função objeto, em matemática, é uma função que precisa ser minimizada ou maximizada usando métodos numéricos para equações não lineares, pois não tem solução analítica.

No *PMF* a função objeto Q é calculada conforme a equação 3.2, onde u_{ij} é encontrado isolando-o na equação 3.1.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{e_{ij}}{u_{ij}} \right]^2 \tag{3.2}$$

O *PMF* minimiza a função objeto Q e converge quando encontra um Q mínimo local ou global.

Como sempre estamos interessados no mínimo global, precisa-se verificar se a solução gerou um Q mínimo local ou global. A estratégica para identificar o tipo de mínimo é rodar o *PMF* para diferentes valores de *Random Seed* e acompanhar a estabilidade de Q.

A equação 3.2 foi inicialmente implementada usando o *método de Gauss-Newton* (citar fonte 1997 Paatero Tapper), mas na versão atual do software *PMF EPA 5.0* é usado o *Método do Gradiente Conjugado*, que necessita de menos recursos computacionais.

Os valores de g_{ik} e f_{kj} são ajustados até encontrar menor Q. O *PMF EPA 5.0* nos oferece dois valores para Q: $Q_{verdadeiro}$ e $Q_{robusto}$, onde o primeiro foi calculado considerando todos os valores de concentração e no segundo remove-se os *outliers*. Assim quando há poucos *outliers*, $Q_{verdadeiro}$ e $Q_{robusto}$ são próximos. Incertezas muitos altas também causam $Q_{verdadeiro}$ e $Q_{robusto}$ similares.

O *PMF EPA 5.0* começa com os valores da matriz f_{kj} randômicos, que são então modificados, até a encontrar a melhor solução, ou seja, a do menor $Q_{robusto}$. Devido ao inicio randômico é recomendado uma rodada de pelo menos 100 iterações como solução final, escolhendo-se a do menor $Q_{robusto}$, para termos assim mais esperança de encontrarmos um Q mínimo global e não local.

3.2.4 Pré-processamento do ajuste

Requisitos conceituais e técnicos para realização do ajuste.

Signal Noise (S/N)

Dependendo do método de medida, conhecimento dos processos químicos atmosféricos e limite de detecção é necessário diminuir o peso de algumas espécies no modelo, diminuindo as incertezas. Além do conhecimento da espécie, o *Signal Noise (S/N)* é um bom indicador de quais amostra deve-se aumentar as incerteza.

- Se, $c_{ij} > u_{ij}$, então $S/N = (c_{ij} - u_{ij})/u_{ij}$.
- Se, $c_{ij} < u_{ij}$, então $S/N = 0$.

O S/N indica se a variabilidade nas medidas é real ou faz parte do ruído dos dados. Espécie com a concentração menor que a incerteza, não apresentam ruído, e devem ser removidas da análise. Espécies com valores muito próximos da incerteza, portanto com S/N próximo de zero ($< 0{,}5$) também devem ser removidas da análise. Espécie nas quais a concentração é pelo menos duas vezes o valor da incerteza, isto é, S/N é maior ou igual à 1, são as ideais para análise, e não altera-se as incertezas.

Quando S/N está entre 0,5 e 1, diminuímos o peso da espécie aumentando a incerteza em um fator 3.

Pré-requisitos do *Software*

Os dados de concentração, c_{ij}, e incertezas, u_{ij}, devem seguir alguns pré-requisitos antes da realização do ajuste:

- Não é permitido concentrações negativas em c_{ij} ou em u_{ij}.
- células vazias não são aceitas.
- nomes das colunas(espécies) e linhas(amostras) devem ser únicos.
- é ideal (não obrigatório) que os dados já estejam classificados pela data em ordem crescente.

Inspeção dos dados de entrada

É necessário fazer ainda inspeções ou alterações antes de fazer o ajuste.

- Espera-se uma relação tipicamente linear entre **Concentração** × **Incerteza**. Investigar o motivo caso isso não ocorra.

- Adequação das incertezas baseado no *Signal Noise (S/N)* e conhecimento da espécie.
- Marcar a Massa Total como *Total Variable*
- Inspecionar gráficos de dispersão entre espécies nas quais se espera correlação, anti-correlação ou não correlação.
- Inspecionar gráficos de séries temporais, para:
  - Identificar padrões temporais em espécies individuais ou em grupo de espécies.
  - Remover *outliers*

3.2.5 Fazendo o ajuste

Segue-se uma séries de conceitos que deverão ser entendido antes da realização do ajuste.

Ambiguidade Rotacional

A comparação fator por fator da contribuição dos fatores, g_{ik}, pode ser plotada em um tipo de gráfico chamado *G-Space*. O gráfico *G-Space* auxilia na verificação da existência de *Ambiguidade Rotacional* na solução. Dada a natureza da *Análise Fatorial*, onde os fatores não devem estar correlacionados, no gráfico *G-Space* fatores com pontos fora da proximidades dos eixos tem maior *Ambiguidade Rotacional*.

Amostras com contribuição zero em ambos os eixos proporcionam maior estabilidade na solução e portanto menor *Ambiguidade Rotacional*.

Checklist

Uma vez que apresentamos os conceitos necessários o entendimento do jargão usado no método *PMF*, apresentamos as etapas escolhidas para realização do ajuste.

1. Rodar o *base run* com 20 iterações, encolhendo o número de fatores variando de 3 até 10. Verificar a quantidade de fatores com melhor significado físico. As matrizes soluções g_{ik} e f_{kj} sairão nos arquivos: *profiles.csv* e *contributions.csv*.
2. Verificar a estabilidade do $Q_{verdadeiro}$ e $Q_{robusto}$ que convergiram, se Q não for estável, então não foi um bom ajuste.

3. Regressão linear simples das concentrações das espécies ajustadas versus medidas, que devem estar correlacionas. Remover amostras ou aumentar as incertezas das amostras que não foram bem ajustadas. Rodar novamente o *base run*.

4. Série temporal das concentrações das espécies ajustadas sobreposta as medidas. Identificar pontos não bem ajustados, devido a fontes infrequentes, por exemplo. Removê-los ou aumentar a incerteza. Rodar novamente o *base run*.

5. Análise residual. Verificar se a distribuição do resíduo é normal (usando o *Teste de Kolmogorov–Smirnov*). Quando não é normal há a indicação que o ajusto foi pobre para essa espécie. Pode-se diminuir o peso da espécie na análise aumentando sua incerteza, ou até mesmo remover a espécie da análise.

6. Verificação se Q é mínimo global ou Local usando 10 valores diferentes de *Random Seed*.

7. Avaliação da *Ambiguidade Rotacional* usando os gráficos *G-Space*.

8. Série temporal de g_{ik}. Avaliação do significado físico.

3.2.6 Pós-processamento do ajuste

Métodos para avaliação da estabilidade da solução.

filtro: PTFE, teflon, 37mm de diametro

PM10 medido com Harvard Impactor com D_50, diamtro aerodinamico: 10 μm] em 5LPM (+/-10 %)

Duas placas impactadoras consecticvas (com óleos) servindo como superficie de impacto.

PM2.5 também usou o Harvard impactor combinado inlet seletivo PM2.5 de com PUF (espuma de polyurethane) com D_50 de 2.5 at 5LPM (+/-10 %)

3.3 Fluorescência de Raios X

Avaliação quali-quantitativa da amostras. Técnica não destrutiva, simultânea. Instrumental: sem pré-tratamento químico.

Tipos de fluorescências de Raios X:

- Dispersão de Energia
- dDispersão de comprimento de onda

- Reflexão total

A dispersão por comprimento de onda é baseada na lei de Bragg. Os dispersivos em energia usam semicondutor capaz de discriminar energia próximas.

Os raios X excitador pode ser um tudo ou fonte radioativa emissoras de raios x. (no nosso caso não seria dedido a baixa inetnsidade)

Correção do efeito matriz: interações do raios x característicos com os elementos da amostras. (absorção do raiox x ou reforço)

fases do edx;

outras técnicas de excitação: partículas aceleradas: elétrons, prótons ou íons, alfa e beta negativa.

Energia de ligação pode ser aproximada pela teoria atomica de bohr.

$E = \frac{me^4(Z-b)^2}{8w^2h^2n^2}$

E = energia de ligação eletrônica (joules), m = massa de repouso do elétron = 9,11.10 -31 kilogramas, e = carga elétrica do elétron = 1,6.10 -19 coulombs, Z = número atômico do elemento emissor dos raios X, b = constante de Moseley, com valores iguais a 1 e 7,4, para as camadas K e L, respectivamente. w = o = permitividade elétrica no vácuo = 8,8534.10 -12 coulombs.newton -l .metro -2 , h = constante de Planck = 6,625.10 -34 joules.s, e n = n o quântico principal do nível eletrônico (n = 1 para camada K, n = 2 para camada L, etc.),

amplitude do pulso eletrônico produzido no detector.

bragg: critais de difração com distancias interplanares conhecidas

kalfa, kbeta: representa transições. kalfa: L para K. Kbeta: M para K

Fazer diagrama das transições .Lembrando que há transições proíbidas. Algumas transições tem energias tão próximas que é impossível seprara.

rendimento: raios-x efetivamente emitidos em relação as vacancias produzidas.

espectrômetro de raios X por dispersão de energia

pulsos ele-trônicos proporcionais às energias dos raios X

o mais empregado é o detector de silício ativado com lítio, Si(Li),

3.4 Limite de Detecção

Bibliografia

Arku, R. E., Vallarino, J., Dionisio, K. L., Willis, R., Choi, H., Wilson, J. G., Hemphill, C., Agyei-Mensah, S., Spengler, J. D., and Ezzati, M. (2008). Characterizing air pollution in two low-income neighborhoods in accra, ghana. *Science of the total environment*, 402(2):217–231.

Dionisio, K. L., Rooney, M. S., Arku, R. E., Friedman, A. B., Hughes, A. F., Vallarino, J., Agyei-Mensah, S., Spengler, J. D., and Ezzati, M. (2010). Within-neighborhood patterns and sources of particle pollution: mobile monitoring and geographic information system analysis in four communities in accra, ghana. *Environmental health perspectives*, 118(5):607–613.

Ezzati, M., Lopez, A. D., Rodgers, A., and Murray, C. J. (2004). Comparative quantification of health risks. *Global and regional burden of disease attributable to selected major risk factors. Geneva: World Health Organization.*

Montgomery, M. R. (2008). The urban transformation of the developing world. *Science*, 319(5864):761–764.

Norris, G., Duvall, R., Brown, S., and Bai, S. (2014). EPA POSITIVE MATRIX FACTORIZATION (PMF) 5.0 Fundamentals & User Guide. *Prepared for the US Environmental Protection Agency, Washington, DC, by the National Exposure Research Laboratory, Research Triangle Park.*

Riley, L. W., Ko, A. I., Unger, A., and Reis, M. G. (2007). Slum health: diseases of neglected populations. *BMC international health and human rights*, 7(1):2.

Sclar, E. D., Garau, P., and Carolini, G. (2005). The 21st century health challenge of slums and cities. *The Lancet*, 365(9462):901–903.

Smith, K. R., Mehta, S., and Maeusezahl-Feuz, M. (2004). Indoor air pollution from household use of solid fuels. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, 2:1435–93.