Estatística Multivariada Aplicada ao Estudo da Qualidade do Ar na cidade do Rio de Janeiro

Roberto Campos Leoni
rcleoni@yahoo.com.br
AEDB e AMAN

Nilo Antonio de Souza Sampaio
nilo.sampaio@uerj.br
UERJ - FAT e AEDB

Sérgio Machado Corrêa
sergio@air.pro.br
UERJ - FAT

Resumo:O estudo da qualidade do ar de uma cidade envolve diferentes aspectos como as emissões, as transformações físico químicas, a meteorologia e a topografia, tornando o estudo de qualidade do ar altamente não linear. Em particular as cidades brasileiras experimentam um mosaico bem distinto do panorama mundial, em função do uso de diferentes combustíveis utilizados por nossa frota veicular. Neste estudo, utilizou-se um conjunto de dados coletados por uma estação de monitoramento da qualidade do ar na cidade do Rio de Janeiro e estes dados foram tratados estatisticamente, de forma descritiva e multivariada, empregando-se análise dos componentes principais e agrupamentos euclidianos e critérios de Ward. Os resultados indicaram que há correlação entre os poluentes primários monóxido de nitrogênio e monóxido de carbono, sugerindo que possuem a mesma origem veicular e que o monóxido de nitrogênio a radiação solar e o ozônio também estão correlacionados, corroborando a formação fotoquímica do ozônio. Outras conclusões são igualmente interessantes, tais como a temperatura e a umidade relativa do ar são inversamente correlacionadas e que o ozônio possui uma contribuição de localidades vizinhas, em função da dependência deste com a velocidade do vento.

Palavras Chave: qualidade do ar - multivariada - dendograma - mapa fatorial -

1. INTRODUÇÃO

A qualidade do ar de uma grande cidade é o resultado das emissões atmosféricas das fontes antropogênicas fixas e móveis (veicular), das fontes naturais, dos processos de deposição via seca e úmida, do transporte entre localidades, da topografia e das transformações físico químicas que ocorrem na atmosfera, com a conversão dos poluentes primários emitidos em poluentes secundários, como é o caso do ozônio (Orlando et al., 2010).

O estudo de qualidade do ar é um tema complexo de pode ser abordado de diferentes modos. Porém, os resultados são de difícil interpretação pois a química da atmosfera é altamente não linear (Martins et al., 2015; Teixeira et al., 2012). Os modelos de qualidade do ar e de previsão podem ser classificados de diferentes modos (Lora, 2002). Pode-se classificar um modelo por sua estrutura básica em determinísticos ou estocásticos ou estacionário dependente do tempo. Outra classificação é por seu marco de referência, como um modelo Euleriano ou Lagrangiano. Com relação à sua dimensionalidade, tem-se os modelos adimensionais, unidimensionais, bidimensionais ou tridimensionais. Finalmente, pelo método de resolução das suas equações fundamentais, tem-se os modelos analíticos ou numéricos.

A química da atmosfera urbana tem seu foco nos compostos orgânicos voláteis (COV) e suas reações fotoquímicas envolvendo os óxidos de nitrogênio ($NOx = NO + NO_2$), com a consequente formação do ozônio (O_3) troposférico e uma série de COV oxidados (Finlayson-Pitts e Pitts, 2000). Nestas reações o radical hidroxila (•OH) reage rapidamente com os COV antropogênicos formando radicais intermediários $\bullet RO_2$ e $\bullet HO_2$, que reagem com o NO convertendo-o a NO_2, que decompõe fotoquimicamente a um átomo de oxigênio excitado $O(^3P)$ e NO. A reação do $O(^3P)$ com o oxigênio molecular (O_2) é a única fonte antropogênica do ozônio na troposfera (Atkinson, 2000). A decomposição do ozônio forma outro átomo de oxigênio excitado $O(^1D)$ que reage com o vapor d'água formando os radicais •OH que reiniciam todo o processo.

Neste trabalho pretende-se abordar a qualidade do ar analisando-se a interdependência dos dados coletados em uma estação automática da qualidade do ar na cidade do Rio de Janeiro através de técnicas estatísticas multivariadas e não sob a ótica de modelos de qualidade do ar.

As demais seções do trabalho apresentam os procedimentos metodológicos, os resultados e a discussão acerca da qualidade do ar, algumas considerações finais e perspectivas de trabalhos futuros.

2. METODOLOGIA

Os dados foram coletados por uma estação móvel de monitoramento da qualidade do ar da Secretaria Municipal de Meio Ambiente (SMAC) da Prefeitura da Cidade do Rio de Janeiro, localizada na PUC-Rio, entre os meses de julho a outubro de 2011, compreendendo as estações do inverno e primavera. Estava localizada na Rua Marquês de São Vicente, 225, Gávea, (22° 58' 44" S e 43° 13' 54" W, altitude 20 m), em um estacionamento, o que ocasiona uma atmosfera estagnada ocasionada pela concentração de emissões evaporativas advindas dos veículos estacionados. Além disso, a PUC-Rio está próxima à Lagoa Rodrigo de Freitas e do mar, o que proporciona altos valores de velocidade do vento e umidade. Cabe ressaltar que também está presente nas proximidades da PUC-Rio uma das vias trânsito mais movimentadas da cidade, a Auto Estrada Lagoa – Barra. Segundo o site da Prefeitura do Rio de Janeiro, esta via possui um fluxo diário de aproximadamente 130 mil veículos. De acordo com a localização podem ser esperados ventos mais concentrados de poluentes vindos, de forma decrescente, do Jardim Botânico e Botafogo, da auto estrada, da montanha (Rocinha), do Leblon, do mar e da floresta da Tijuca. Sendo assim, esta região sofre influência de diferentes tipos de fontes de emissão (Luna et al., 2014).

As variáveis consideradas neste estudo foram: dióxido de nitrogênio (NO_2), monóxido de nitrogênio (NO), óxidos de nitrogênio (NOx), monóxido de carbono (CO), ozônio (O_3), velocidade escalar do vento (VEV), radiação solar global (RSG), temperatura (TEM), umidade relativa (UR) e precipitação pluviométrica (PP). A variável PP foi utilizada para determinar a retirada dos dados dos dias chuvosos e quando o seu valor se apresentava diferente de zero, estes dados eram expurgados do banco, pois a chuva reduz o nível dos poluentes atmosféricos por deposição úmida. Além disso, os dados obtidos nos sábados, domingos e feriados foram também desconsiderados, pelo reduzido tráfego veicular. Os dados noturnos também foram removidos, pois à noite a camada de mistura da atmosfera é muito baixa, o que aumenta a concentração de alguns poluentes primários e também a reação fotoquímica é praticamente inexistente, produzindo níveis de ozônio muito reduzidos. O estudo da atmosfera urbana noturna deve ser um estudo à parte.

Para as medições do CO foi utilizado o analisador Ecotech modelo EC9830, que realiza medidas na faixa de 0 a 200 ppm com limite de detecção de 50 ppb. Para as medições dos NOx foi utilizado o analisador Ecotech modelo EC9841, com medidas na faixa de 0 a 20 ppm com limite de detecção de 0,4 ppb. Para as medições do O_3 foi utilizado o analisador da Ecotech modelo EC9810 que opera na faixa de 0 a 20 ppm com limite de detecção de 0,5 ppm.

Em Estatística Multivariada a análise de agrupamentos representa um conjunto de técnicas exploratórias muito úteis e que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos semelhantes entre observações em relação a determinadas variáveis e o objetivo de se criarem grupos, ou clusters, em que prevaleça a homogeneidade interna. Nesse sentido, esse conjunto de técnicas, também conhecido por análise de conglomerados ou análise de clusters, tem por objetivo principal a alocação de observações em uma quantidade relativamente pequena de agrupamentos homogêneos internamente e heterogêneos entre si e que representam o comportamento conjunto das observações a partir de determinadas variáveis (Fávero e Belfiore, 2015).

A extração de informações dos resultados de um experimento químico envolve a análise de grande número de variáveis. Muitas vezes, um pequeno número destas variáveis contém as informações químicas mais relevantes, enquanto que a maioria das variáveis adiciona pouco ou nada à interpretação dos resultados em termos químicos. A decisão sobre quais variáveis são importantes é feita, geralmente, com base na experiência, ou seja, baseado em critérios que são mais subjetivos que objetivos. A redução de variáveis através de critérios objetivos, permitindo a construção de gráficos bidimensionais contendo maior informação estatística, pode ser conseguida através da análise de componentes principais. Também é possível construir agrupamentos entre as amostras de acordo com suas similaridades, utilizando todas as variáveis disponíveis, e representá-los de maneira bidimensional através de um dendograma. A análise de componentes principais e de agrupamento hierárquico são técnicas de estatística multivariada complementares que têm grande aceitação na análise de dados químicos (Neto e Moita, 1997).

Duas técnicas estatísticas multivariadas foram combinadas na condução das análises: métodos de componentes principais e agrupamento hierárquico para descrever e visualizar a semelhança entre as variáveis. As técnicas foram implementadas com o pacote *FactoMineR* (Lê et al., 2008) disponível no software R (R Core Team, 2013). Após a seleção dos componentes principais, o agrupamento hierárquico foi realizado com base na medida de distância euclidiana e o critério aglomeração de Ward, pois buscou-se gerar grupos (clusters) que possuam uma alta homogeneidade interna.

3. RESULTADOS E DISCUSSÃO

As medidas descritivas média, desvio padrão e coeficiente de variação das variáveis são apresentadas na Tabela 1. Observa-se que o coeficiente de variação da variável CO é o mais alto quando comparado com as demais variáveis, indicando alta dispersão relativa (136%), seguida por RSG (87%) e NO (82%).

Tabela 1: Estatísticas descritiva dos dados.

Variáveis	Média	Desvio Padrão	Coeficiente de Variação
NO_2	21,89 µg/m3	11,27 µg/m3	51 %
NO	43,03 µg/m3	35,16 µg/m3	82 %
NOx	65,11 µg/m3	39,87 µg/m3	61 %
CO	0,14 ppm	0,19 ppm	136 %
O_3	40,70 µg/m3	25,48 µg/m3	63 %
VEV	1,08 m/s	0,51 m/s	47 %
RSG	311,7 W/m^2	269,8 W/m^2	87 %
TEM	23,25 ºC	3,16 ºC	14 %
UR	69,13 %	15,10 %	22 %

Pode-se explicar a alta variabilidade do CO e do NO por estas serem as espécies inorgânicas predominantemente emitidas pelas fontes móveis, juntamente com os COV, que não foram mensurados no conjunto de dados estudado. A variabilidade da RSG é explicada pela diferença de intensidade do sol ao longo do dia, onde nas primeiras horas da manhã e nas horas finais da tarde são observados períodos de sombra.

A Figura 1 ilustra as correlações entre as variáveis. As variáveis NOx e NO apresentam alto grau de correlação linear positiva (r = 0,95), justificado pelo fato do NOx ser a soma do NO e NO_2, ao passo que as variáveis NOx com RSG e TEM praticamente são não correlacionadas. Altas correlações justificam o uso dos componentes principais na redução das dimensões. Entretanto, variáveis não correlacionadas aos pares inviabilizam o uso dos componentes principais, tornando-se inapropriados para redução da dimensionalidade. Conforme discutem Hair et al. (2009) e segundo Fávero e Belfiore (2015), embora a inspeção visual da matriz de correlações não revele se a extração de fatores será, de fato, adequada, uma quantidade substancial de valores inferiores a 0,30 representa um preliminar indício de que a análise fatorial poderá ser inapropriada. Para que seja verificada a adequação global propriamente dita da extração dos fatores, deve-se recorrer ao teste de esfericidade de Bartlett (1950). Esse teste foi empregado com a finalidade de avaliar a hipótese nula de esfericidade. A rejeição da hipótese nula de esfericidade indica que é apropriado reduzir a dimensionalidade dos dados. A estatística teste de Bartlett calculada com base nas variáveis conduziu os resultados à rejeição da hipótese nula de esfericidade $\left(\chi^2_{36} = 6288{,}2 \;\; \text{p} - \text{valor} \cong 0\right)$, ou seja, justifica-se o uso de componentes principais para reduzir a dimensão.

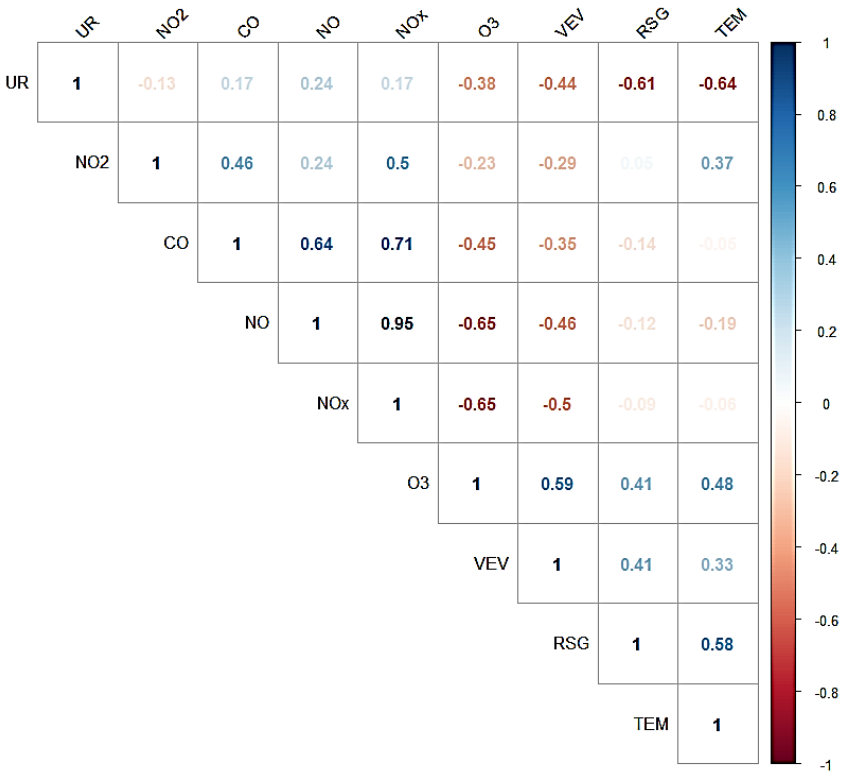

Figura 1: Matrix de correlações entre as variáveis estudadas.

Os autovalores e um resumo das porcentagens da variância total explicada pelas componentes principais são apresentados na Tabela 2. Observa-se que os três primeiros componentes (CP1, CP2 e CP3) acumulam aproximadamente 79% da variabilidade total dos dados. Reter os três primeiros componentes para análises posteriores é bastante razoável para uma representação parcimoniosa das variáveis. A escolha de três componentes parece bom tomando como referência a Tabela 4, verificando-se que, com três componentes principais, a menor porcentagem de variância individual explicada é a da variável VEV com 67%.

A interpretação das componentes será feita com base nas Tabelas 3, 4 e 6 e Figuras 2, 3 e 4. A Tabela 3 apresenta os coeficientes das componentes principais. As variáveis NO, NOx, CO, O_3 e VEV possuem os maiores coeficientes na primeira componente principal, resultado já esperado, pois as variáveis foram padronizadas e as correlações entre as variáveis são altas e algumas próximas umas das outras. A segunda componente principal apresenta os maiores coeficientes para as variáveis NO_2, RSG, TEM e UR.

Tabela 2: Autovalores e explicação da variância total

Componentes	Autovalor	Explicação (%)	Explicação acumulada
CP1	4,02	44,68	44,68
CP2	2,28	25,30	69,98
CP3	0,85	9,45	79,43
CP4	0,57	6,31	85,74
CP5	0,43	4,81	90,56
CP6	0,33	3,64	94,20
CP7	0,32	3,51	97,71
CP8	0,20	2,21	99,92
CP9	0,01	0,08	100,00

Com auxílio da Tabela 4, é possível criar os mapas fatoriais apresentados nas Figuras 2, 3 e 4. A Tabela 6 ilustra as contribuições de cada variável para a componente principal.

Tabela 3: Coeficientes das componentes principais

Variável	CP1	CP2	CP3
NO_2	0,18	-0,43	0,64
NO	0,41	-0,20	-0,45
NOx	0,42	-0,31	-0,21
CO	0,35	-0,27	-0,04
O_3	-0,42	-0,05	0,09
VEV	-0,37	-0,07	-0,35
RSG	-0,25	-0,42	-0,39
TEM	-0,23	-0,50	0,22
UR	0,27	0,42	0,11

Vetores que representam pontos variáveis de alta contribuição, ou seja, extremos próximos à circunferência do círculo de correlações do mapa fatorial, representam as variáveis que justificam a maior dispersão. São essas as variáveis que desempenham um papel mais relevante na análise, pois são as variáveis determinantes da componente principal.

Tabela 4: Correlações entre as variáveis e as componentes principais

Variável	CP1	CP2	CP3
NO_2	0,36	-0,65	0,59
NO	0,83	-0,30	-0,41
NOx	0,84	-0,46	-0,19
CO	0,70	-0,40	-0,03
O_3	-0,84	-0,08	0,08
VEV	-0,74	-0,10	-0,33
RSG	-0,50	-0,63	-0,36
TEM	-0,46	-0,76	0,20
UR	0,55	0,63	0,10

Tabela 5: Porcentagem explicada das variâncias individuais

Variável	CP1	CP2	CP3	Soma
NO_2	0,13	0,42	0,35	0,90
NO	0,68	0,09	0,17	0,94
NOx	0,71	0,21	0,04	0,96
CO	0,49	0,16	0,00	0,65
O_3	0,71	0,01	0,01	0,73
VEV	0,55	0,01	0,11	0,67
RSG	0,25	0,39	0,13	0,77
TEM	0,21	0,58	0,04	0,83
UR	0,30	0,40	0,01	0,71

Tabela 6: Contribuição das variáveis nas componentes principais

Variável	CP1	CP2	CP3
NO_2	3,21	18,53	41,35
NO	16,95	3,98	19,94
NOx	17,57	9,31	4,22
CO	12,13	7,13	0,14

O_3	17,64	0,25	0,80
VEV	13,60	0,47	12,50
RSG	6,12	17,26	14,86
TEM	5,31	25,40	4,93
UR	7,47	17,68	1,27

A análise do mapa fatorial apresentado na Figura 2 evidencia que as variáveis NO, NOx, CO e NO_2 encontram-se positivamente correlacionadas, essa última de forma mais fraca em relação às outras três, pois sua representação no mapa não fica próxima ao círculo das correlações. Esta correlação entre as variáveis NO, NO_2, NOx e CO indica que estes poluentes possivelmente são oriundos das mesmas fontes, que no caso da localidade estudada são as fontes veiculares.

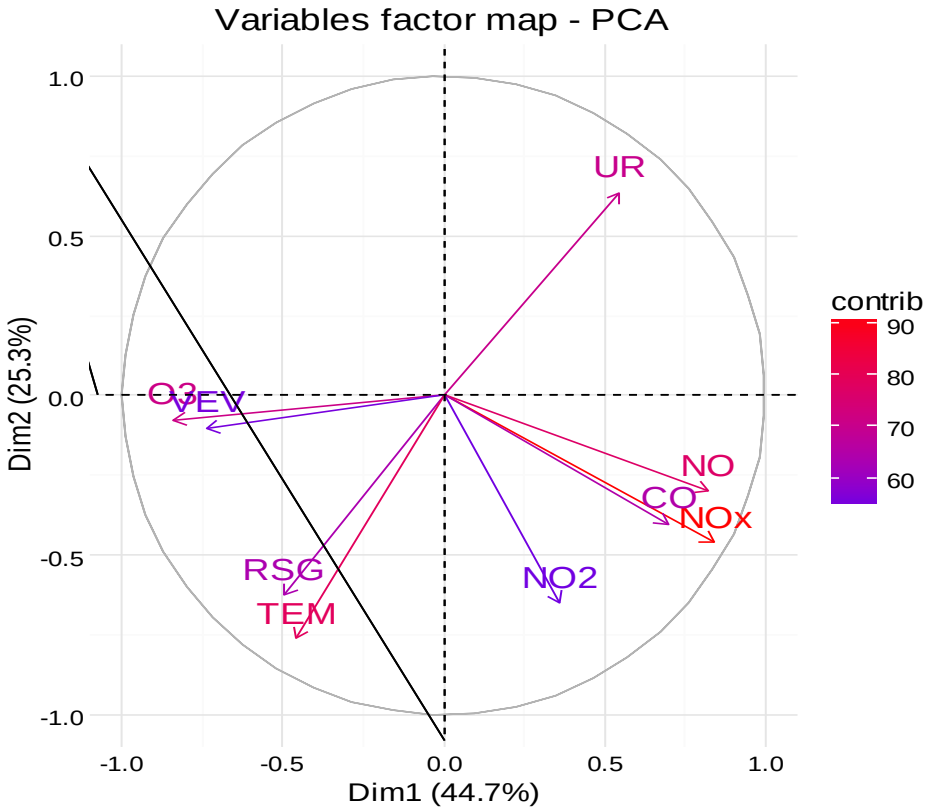

Figura 2: Mapa Fatorial (CP1 x CP2).

Ozônio e VEV também são positivamente correlacionadas, mas as desse grupo apresentam correlações negativas com as do grupo anterior; RSG e TEM são positivamente correlacionadas e não apresentam correlação significante com as do grupo (NO, NOx, CO e NO_2), devido às quase ortogonalidades entre os vetores. Pode-se explicar que a alta correlação entre RSG e TEM deve-se ao fato que as duas são oriundas do sol. Uma correlação positiva entre VEV e O_3 pode indicar que este poluente pode estar sendo transportado de localidades vizinhas, sendo uma parte produzido localmente e outra parte produzido em regiões de maior produção de NOx e COV, seus precursores. Também é possível observar uma forte correlação entre O_3 e TEM, visto que o ozônio é formado por processos fotoquímicos dependentes da luz do sol e da temperatura. Outra observação oriunda da Figura 2 é a correlação inversa entre TEM e UR. No início da manha costuma-se observar baixos valores de TEM e altos valores de UR. Com o aumento da TEM ao longo do dia a água passa para a fase gasosa reduzindo o valor da UR e incrementando o valor da pressão atmosférica. No final do dia comportamento similar é

observado, com a redução da TEM e incremento da UR. Análise semelhante pode ser feita nos mapas das Figuras 3 e 4.

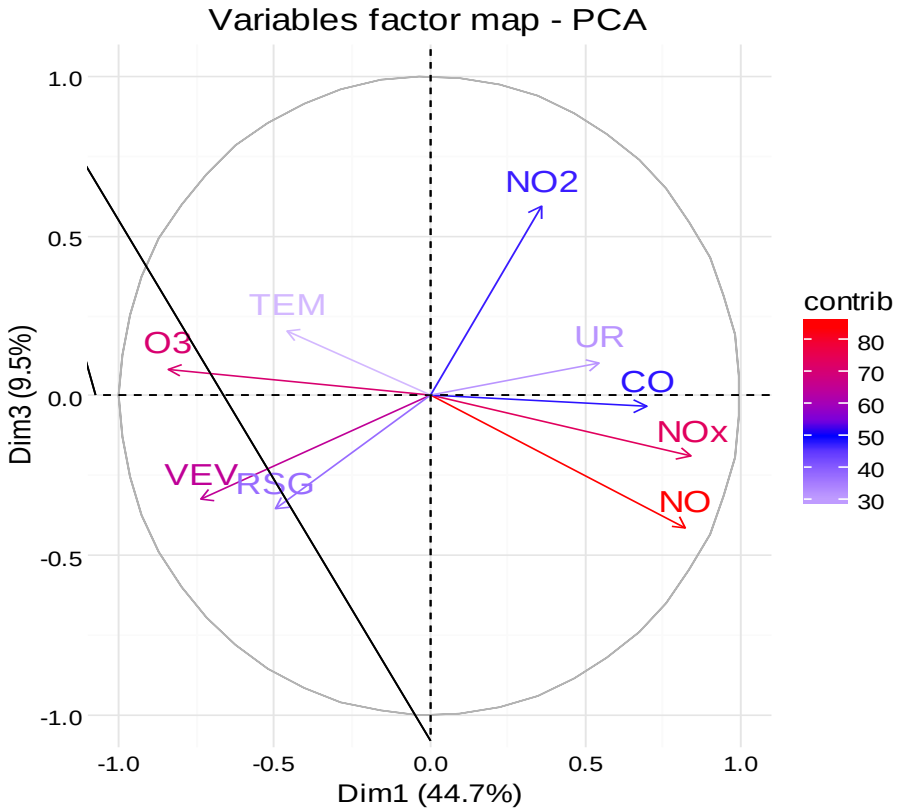

Figura 3: Mapa Fatorial (CP1 x CP3).

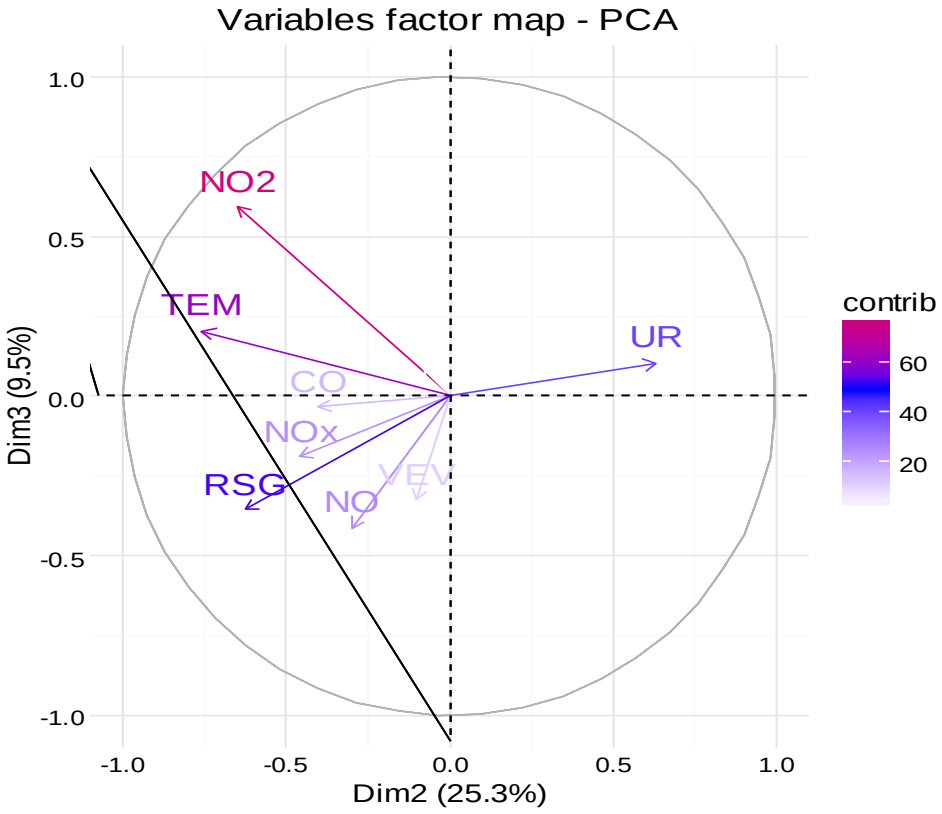

Figura 4: Mapa Fatorial (CP2 x CP3).

Com base nas componentes principais extraídas, adotando-se a medida de distância euclidiana e o critério de Ward, procurou-se alocar as variáveis em uma quantidade de agrupamentos (*clusters*) homogêneos internamente e heterogêneos entre si. Na Figura 5 é apresentado um dendograma com as distancias euclidianas com os clusters propostos, que ilustram um resumo dos grupos formados. A variável UR foi alocada em um cluster individual; as variáveis (NO, NOx, CO e NO_2) constituem um segundo *cluster* e as demais variáveis (O_3, VEV, RSG, TEM) formam o terceiro *cluster*.

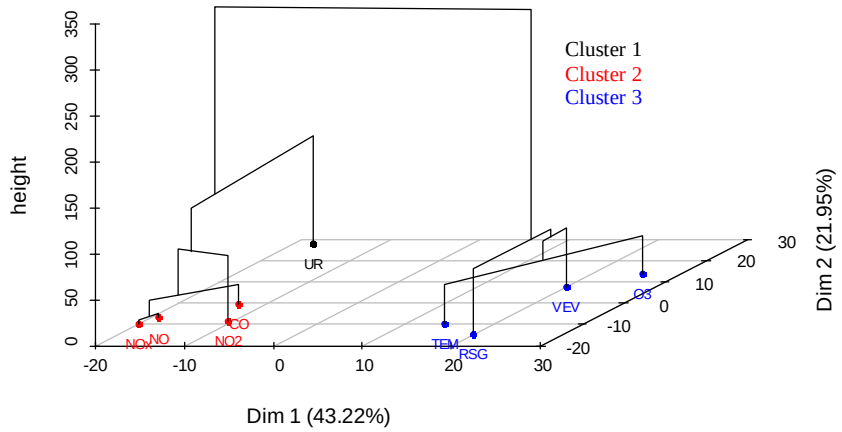

Figura 6: Dendograma das variáveis estudadas separadas por clusters.

4. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Este estudo permitiu correlacionar as variáveis mensuradas por uma estação automática de monitoramento da qualidade do ar na cidade do Rio de Janeiro e desta forma auxiliar no entendimento da inter-relação das variáveis, em uma sinergia entre a química da atmosfera e as ferramentas estatísticas.

Foi possível entender que as fontes veiculares são prioritárias na região estudada, com alta correlação entre os poluentes primários CO e NO, emitidos diretamente pelas fontes móveis. Apesar dos COV não terem sido monitorados neste estudo o ozônio também apresentou uma correlação com o NOx, um dos precursores do ozônio, assim como a correlação deste com a radiação solar e temperatura. O ozônio também apresentou uma correlação com a velocidade do vento, o que pode indicar que está sendo transportado de outras localidades.

Estudos estão em desenvolvimento para outras localidades da cidade para um banco de dados mais robusto, que envolve diferentes anos e estações do ano, incluindo também os COV de uma forma global e alguns outros de forma especiada, para desta forma poder tentar chegar a um conjunto mínimo de dados a serem monitorados e conhecer a qualidade do ar.

Espera-se que este estudo e vindouros venham auxiliar as agências ambientais a tomarem decisões sobre a qualidade do ar e mesmo chegar a um conjunto mínimo de parâmetros a serem monitorados, reduzindo desta forma os elevados custos de se instalar e operar uma estação automática da qualidade do ar, que em geral são de 100 a 300 de milhares de dólares anuais.

5. REFERÊNCIAS

Atkinson, R. Atmospheric chemistry of VOCs and NOx. Atmospheric Environment, 34, 2063-2101, 2000.

Bartlett, M.S. Tests of Significance in Factor Analysis. British Journal of Statistical Psychology, 3, 77–85, 1950.

Fávero, L.P; Belfiore, P. Análise de Dados: Técnicas Multivariadas Exploratórias com SPSS e STATA. Elsevier Editora, 2015.

Finlayson-Pitts, B. J.; Pitts, J. N. Chemistry of the Upper and Lower Atmosphere. San Diego: Academic Press, 2000.

Hair J.R.; Black J.F.; Babin, W.C.; Anderson, B.J.; Tatham, R.E. Análise multivariada de dados. 6 ed. Porto Alegre: Bookman, 2009.

Lê, S.; Josse, J.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25, 1-18, 2008.

Lora, E.E.S. Prevenção e controle da poluição nos setores energéticos, industrial e de transportes: Interciência, 2002.

Luna, A.S.; Paredes, M.L.L.; Oliveira, G.C.G.; Corrêa, S.M. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. Atmospheric Environment 98, 98-104, 2014.

Martins, E.M.; Nunes, A.C.L.; Corrêa, S.M. Understanding Ozone Concentrations During Weekdays and Weekends in the Urban Area of the City of Rio de Janeiro. Journal of the Brazilian Chemical Society 26, 1967-1975, 2015.

Neto, J.M.M; Moita, G.C. Uma introdução à análise exploratória de dados multivariados. Scielo Editora, 1997.

Orlando, J.P.; Alvim, D.S.; Yamazaki, A.; Corrêa, S.M.; Gatti, L.V. Ozone precursors for the São Paulo Metropolitan Area. Science of the Total Environment, 408, 1612-1620, 2010.

R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Teixeira, J.R.; Souza, C.V.; Sodré, E.D.; Corrêa, S.M. Volatile Organic Compound Emissions from a Landfill, Plume Dispersion and the Tropospheric Ozone Modeling. Journal of the Brazilian Chemical Society, 23, 496-504, 2012.