

USO DE ANÁLISE MULTIVARIADA PARA MAPEAMENTO DO PERFIL DE INTERNACIONALIZAÇÃO DAS UNIVERSIDADES FEDERAIS BRASILEIRAS:

um estudo exploratório a partir de dados disponíveis na base Web of Science

Dalton Lopes Martins (UFG)

dmartins@gmail.com

Eixo temático: Técnicas e Ferramentas para Estudos Bibliométricos e Cientométricos

Modalidade: Apresentação oral

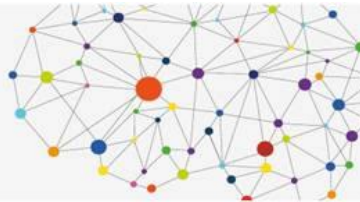
1 INTRODUÇÃO

A importância da internacionalização da produção científica é hoje uma discussão que toca de forma direta e central as diversas estratégias de políticas que visam estimular essa produção e ampliar seu potencial de impacto. A internacionalização da pesquisa amplia não apenas o potencial de visibilidade, logo de referenciamento, de um trabalho acadêmico, como também torna-se estratégico como meio de ampliar o potencial de estabelecimento de parcerias internacionais.

O que impulsiona o desenvolvimento científico e tecnológico são a cooperação e a internacionalização. É um truísmo afirmar que o conhecimento e a inovação têm um papel significativo no desenvolvimento dos diferentes países e, portanto, na melhoria das condições de vida de suas populações. No entanto, como cada sociedade não pode reinventar a roda, é imperioso que haja cooperação entre suas comunidades de pesquisadores. Por isso, os diferentes países têm uma preocupação muito grande com a internacionalização da produção científica (FIORIN, 2007, pag. 264).

Um dos pontos chave para avaliação, portanto, de políticas científicas torna-se mapear o desenvolvimento da internacionalização da produção científica, identificando elementos que influenciem e as diferentes estratégias utilizadas por pesquisadores, instituições bem como seus modos de produção. Para tanto, o uso de bases de dados estruturadas permitem uma maior visibilidade da produção científica (MUGNAINI, LEITE e LETA, 2011), consolidando pontos de acesso que viabilizam a realização de estudos comparativos e análise dos mesmos conjuntos de dados.

Outro ponto chave para essa avaliação é a técnica de análise utilizada em relação aos conjuntos de dados disponíveis para tal. Um dos pontos que temos observado como potencial para o desenvolvimento da área da cientometria no Brasil é a utilização de técnicas estatísticas que permitam levar em consideração múltiplas variáveis que podem descrever o mesmo campo de análise. O interesse em estudar um conjunto de elementos, sejam eles pessoas, instituições ou outros aspectos que venhamos a considerar para nossas análises a partir de um



conjunto múltiplo de variáveis reside no fato de que essas variáveis estão conectadas e podem revelar estruturas de relacionamento que nos permitem compreender melhor o fenômeno estudado (HUSSON, LÊ, PAGÉS, 2011). Não apenas as variáveis informam aspectos interessantes desse campo de análise mas a relação entre essas variáveis torna-se uma fonte de informação relevante para outros possíveis olhares para o campo.

Dessa forma, o objetivo deste estudo é explorar o potencial de uso da análise multivariada de componentes principais a partir de dados coletados na base Web of Science em relação a produção científica das 59 universidades federais.

2 ANÁLISE MULTIVARIADA DE COMPONENTES PRINCIPAIS

A análise de componentes principais se aplica a tabelas de dados onde as linhas são consideradas os indivíduos a serem pesquisados e as colunas as variáveis quantitativas a serem levadas em consideração. Essa análise permite o estudo tanto da relação entre as variáveis quanto da relação entre os indivíduos. Estudar os indivíduos significa identificar similaridades entre eles, ou seja, dado o conjunto de variáveis em questão de que maneira os indivíduos classificados por essas variáveis são próximos ou distantes. Já estudar as variáveis permite identificar os múltiplos relacionamentos lineares que existem entre elas, já que o método de componentes principais apenas lida com esse tipo de relacionamento entre variáveis (HUSSON, LÊ, PAGÉS, 2011).

Um dos aspectos úteis a serem explorados pela análise de componentes principais é a possibilidade de agrupar e reduzir o número de variáveis significativas que deveriam ser utilizadas para análise de um problema. Ou seja, o método apresenta como as diferentes variáveis contribuem para as diferenças ou similaridades entre os indivíduos, de onde podemos chegar a conclusão que duas ou mais variáveis possuem o mesmo comportamento, logo utilizar apenas uma permitiria avaliar a mesma tendência no conjunto de dados.

Dessa forma, o objetivo central da análise de componentes principais é permitir tirarmos conclusões das relações lineares entre as variáveis detectando quais são as dimensões principais responsáveis pela variabilidade no dados (HUSSON, LÊ, PAGÉS, 2011) o que explicaria, portanto, quais são as principais dimensões variáveis que devem ser levadas em consideração na caracterização do conjunto de dados em questão.

Entendemos que esse tipo de análise pode ser utilizado em diversos tipos de problemas de interesse da cientometria e bibliometria, tais como a caracterização de perfis de grupos,

instituições, pesquisadores, artigos, citações, entre outros objetos de análise em relação a diferentes conjuntos de variáveis utilizados para descrever esses objetos. Tanto para a realização de estudos comparativos de indivíduos quando de variáveis utilizadas para descrevê-los, a análise de componentes principais parece contribuir com as pesquisas na área.

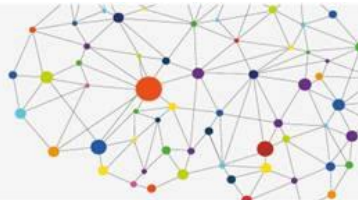
3 METODOLOGIA

Para a realização deste estudo, organizamos a metodologia de pesquisa em 3 etapas, sendo elas, definição e extração de dados, tratamento e enriquecimento dos dados e, por fim, realização da análise e obtenção dos mapas do relacionamento entre variáveis e indivíduos pesquisados.

Na etapa de definição e extração dos dados, inicialmente obtivemos junto ao site do Ministério da Educação¹ o nome das 59 universidades federais a serem pesquisadas, o que representa o total de universidades federais brasileiras. Com isso, montamos uma tabela de base. A partir do nome de cada universidade, realizamos uma pesquisa pelo campo “endereço” na base Web of Science, retornando portanto todos os artigos registrados na base em qualquer revista e área científica em que um pesquisador tenha publicado e mencionado o endereço de sua universidade. Vale dizer que filtramos os dados para apenas os últimos 3 anos (2011-2013), quando na totalidade das universidades já havia ao menos algum material publicado na base de referência. Os dados fornecidos pela base Web of Science para cada universidade são: número de artigos publicados, número de citações recebidas, número de citações recebidas sem autocitação, média de citação por item e índice-H.

Na etapa tratamento e enriquecimento dos dados, realizamos dois procedimentos para facilitar a análise. Inicialmente, modificamos os nomes das universidades por suas siglas, de maneira a facilitar identificar nos gráficos gerados e expostos a seguir. De forma a normalizar o uso do recurso de autocitação pelas universidades, também calculamos uma nova variável, chamada de taxa de autocitação, calculada pela razão entre o número de autocitações realizadas pelos pesquisadores de uma universidade e o número total de citações recebidas pelos seus artigos. Desse modo, incluímos uma nova variável a ser utilizada pela análise multivariada.

¹<http://portal.mec.gov.br/index.php>



Na etapa de realização da análise, utilizamos o software R² de estatística com o pacote FactoMineR, desenvolvido especialmente para a realização de diversos tipos de análise multivariada.

4 RESULTADOS

Na figura 1, a seguir, temos o mapa de relacionamento das variáveis nas duas principais dimensões do espaço de análise. Vale dizer aqui que essas duas dimensões são responsáveis juntas por 81,55% da variabilidade dos dados, ou seja, elas podem ser consideradas as dimensões mais significativas dessa análise. Dito isto, temos que a distribuição e a estrutura de relacionamento das variáveis nos mostram 3 tendências relevantes a serem consideradas: as 4 variáveis artigos, índice-H, número de citações e número de citações sem autocitações são convergentes na mesma direção, ou seja, representam a mesma tendência entre elas, já a taxa de autocitação é praticamente ortogonal a essas 4 variáveis, demonstrando que essa variável mostra outras tendências na estrutura dos dados. O mesmo ocorre com a média de citação por artigo, porém em menor intensidade devido a sua maior proximidade com as outras.

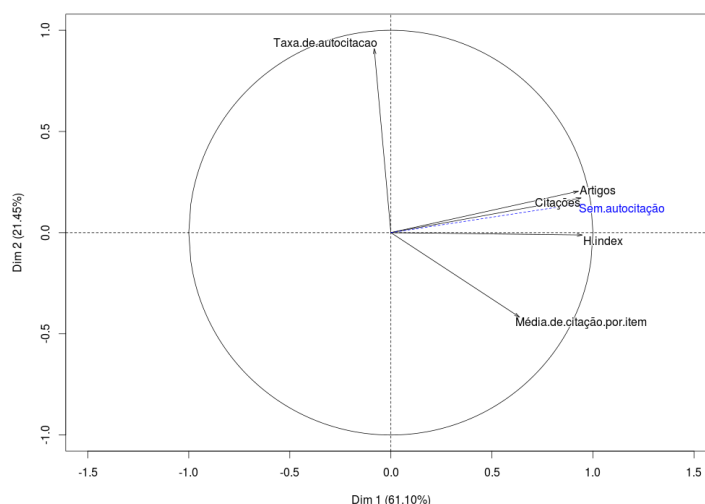
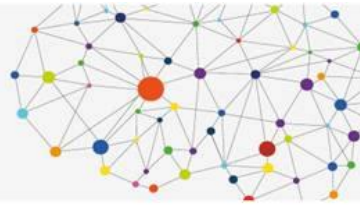


Figura 1. Mapa de variáveis nas duas principais dimensões de análise. Fonte: dados da pesquisa.

De maneira a mostrar um outro modo de visualizar o relacionamento entre essas variáveis, apresentamos na figura 2, a seguir, um mapa de correlação entre elas. Há correlações muito expressivas, como a figura 1 já apresentou, como número de artigos

²<http://www.r-project.org/>



publicados e quantidade de citações recebidas, número de artigos e índice-H, entre outros. O fato que vale ressaltar aqui é a maneira mais sintética e fácil de visualizar que a figura 1 apresenta essas relações, permitindo que ao visualizar a figura 2 possamos entender de maneira pontual as relações de estrutura dos dados entre cada par de variáveis, sem necessariamente se preocupar em apreender da figura a visão do todo sobre o relacionamento entre todas as variáveis.

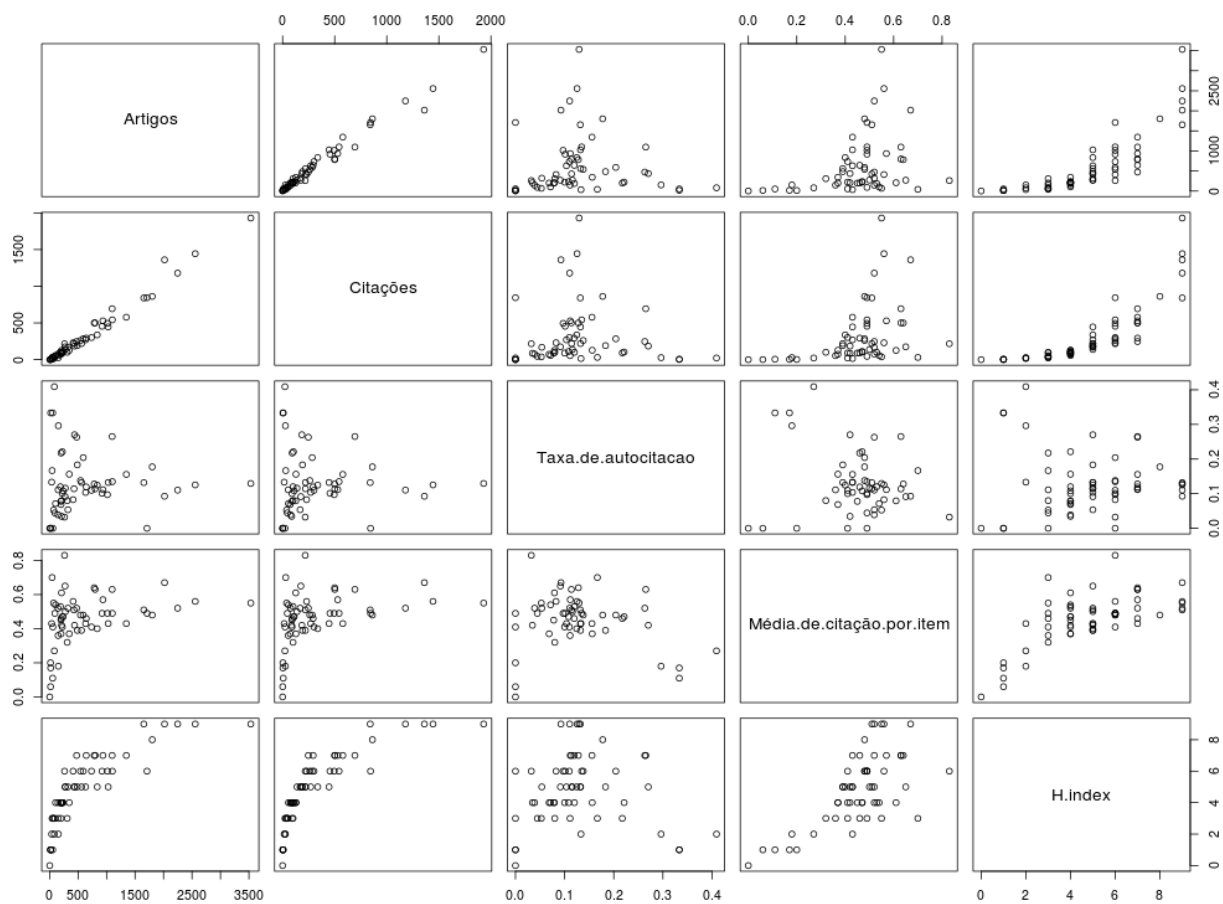
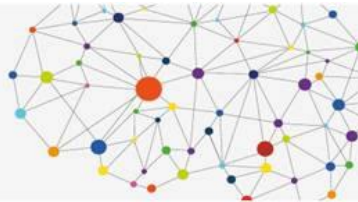


Figura 2. Mapa de correlação entre as variáveis utilizadas. Fonte: dados da pesquisa.

Por fim, apresentamos na figura 3, a seguir, a distribuição das universidades pelas duas dimensões de análise representadas pelas variáveis da figura 1. Ao observar a figura 3, notamos que 5 grandes grupos podem ser percebidos na estrutura de distribuição dos dados. O primeiro deles é representado pela UFI, UFFS, UFAP e UFRAM, sendo esse grupo o de universidades que apresentaram menor quantidade de artigos publicados e citações recebidas, porém com maior taxa de autocitação por parte de seus pesquisadores. O segundo grupo pode ser identificado pela UFSE, UILSA, UFILA, sendo esse grupo das universidades que menos publicaram na população. O terceiro grupo é representado pela UFRG, UFRGS, UFRJ,



UNIFESP, UFSC, UFL, UFV e UNB, sendo o grupo das universidades mais produtivas, com maior quantidade de citações recebidas e índice-H. O quarto grupo representa a maioria dos pontos posicionados próximos ao centro do gráfico, mostrando um perfil próximo entre essas universidades em relação as variáveis utilizadas por este estudo. Já o quinto e último grupo é denotado pela UFCSPA, que se distingue dos demais por ter recebido a maior média de citação por artigo.

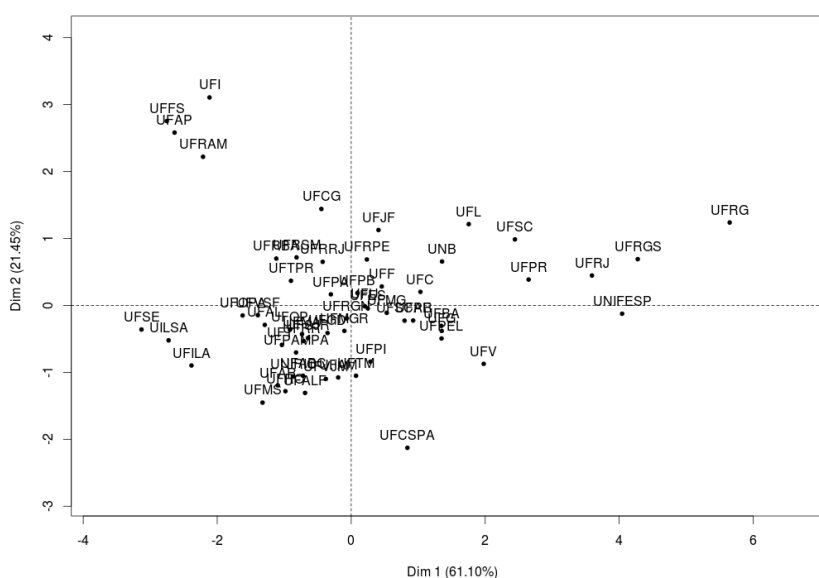


Figura 3. Mapa da distribuição das universidades federais nas duas principais dimensões de análise. Fonte: dados da pesquisa.

5 CONCLUSÃO

A análise multivariada por componentes principais se mostrou neste estudo um método bastante interessante, de fácil uso e implementação para a construção de perfis das universidades federais brasileiras com base nos dados disponíveis sobre sua internacionalização na base Web of Science. Além disso, permitiu identificarmos também de forma simples as principais dimensões de variabilidade dos dados e as formas de relacionamento entre as variáveis utilizadas no estudo, mostrando para estudos futuros que tipos de dados poderiam ser coletados para complementar esse estudo e que tipos de dados seriam redundantes em relação aos já recolhidos.

Entendemos que estudos futuros poderiam explorar outros tipos de análise multivariada, permitindo a categorização das universidades por área geográfica e outros atributos, tais como número de programas de pós-graduação, notas Capes, número de professores, entre outros.

Recomendamos a partir desse estudo a exploração de outros dados e estudos experimentais com análise multivariada nas áreas da cientometria e bibliometria.

REFERÊNCIAS

FIORIN, José Luiz. Internacionalização da produção científica: a publicação de trabalhos de Ciências Humanas e Sociais em periódicos internacionais. **Revista Brasileira de Pós-Graduação**, Brasília, v. 4, n. 8, p. 263-281, dezembro de 2007.

HUSSON, François, LÊ, Sébastien, PAGÉS, Jérôme. **Exploratory Multivariate Analysis by example using R**. CRC Press. 1ed. 2011. 228p

MUGNAINI, Rogério, LEITE, Paula, LETA, Jacqueline. Fontes de informação para análise de internacionalização da produção científica brasileira. **Ponto de Acesso**, Salvador, V.5, n.3 p. 87-102, ago 2011.