em questão
E-ISSN 1808-5245
Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS
v. 20, n.3 – Edição Especial 2014

Uso da 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
Doutor; Universidade Federal de Goiás;
dmartins@gmail.com

Resumo:O artigo tem por objetivo estudar uma técnica de análise multivariada de maneira a contribuir com a discussão e disseminação desse tipo de técnica nos estudos cientométricos no Brasil. Também objetiva aplicar essa técnica e analisar suas possibilidades de aplicação em um estudo de internacionalização da produção científica brasileira com foco nas 59 universidades do sistema federal. O estudo utiliza dados provenientes da base Web of Science, constrói um novo indicador para avaliar a taxa de autocitação das universidades e realiza uma análise multivariada utilizando o software livre estatístico R. Os resultados permitem propor o agrupamento das universidades em cinco grupos devido as suas características reveladas pelas variáveis analisadas. A técnica se mostra de fácil aplicação e muito útil para estudos cientométricos.

Palavras-chave:Análise multivariada. Componentes principais. Internacionalização. Universidades federais. Web of Science.

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 se torna 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, p. 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 esse processo e as diferentes estratégias utilizadas por pesquisadores e instituições, bem como seus modos de produção. Para tanto, o uso de bases de dados estruturadas permite uma maior visibilidade da produção científica (MUGNAINI; LEITE; LETA, 2011), consolidando pontos de acesso que viabilizam a realização de estudos comparativos e a 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.

Leydesdorff (2001, p. 10) descreveu que as características de técnicas que se propõem a estudar fenômenos que podem ser especificados em termos empíricos. O autor explicita que essas técnicas devem tratar de três características principais:

a) ajudar a explicar como cada unidade (caso ou variável) é responsável pela variação de resultados do fenômeno em análise;

b) os possíveis efeitos de agregação e desagregação de variáveis e casos;

c) o modelo de funcionamento dinâmico desse fenômeno.

Colocam-se, portanto, algumas questões de relevância para os estudos cientométricos, de forma geral. Percebe-se a necessidade de sistematização de bases de dados que sirvam de

referência para estudos comparativos e que, ao mesmo tempo, forneçam múltiplas variáveis a respeito de seus objetos de estudo. Esta configuração poderia permitir a aplicação e o exercício de concepção de técnicas analíticas que possibilitem entender como as variáveis se relacionam entre si, como são responsáveis individualmente por explicar a variação dos dados, como podem ser agregadas ou desagregadas na construção de possíveis planos analíticos de interesse para um estudo e, por fim, como podem servir de base de apoio para a representação da dinâmica de operacionalização de um sistema social.

Ao mesmo tempo, sabe-se também que tanto a construção desse tipo de base de dados quanto a aplicação e concepção de técnicas analíticas representam um enorme desafio para o desenvolvimento dos estudos cientométricos no Brasil. Em relação às bases de dados, os custos tecnológicos e operacionais para a constituição de uma massa de dados que represente as principais revistas internacionais nos vários campos do conhecimento e facilite a identificação do posicionamento do Brasil em meio a elas, torna praticamente impeditivo iniciativas nacionais nesse sentido. Já em relação às técnicas analíticas, também se apresenta um desafio. Muitas dessas técnicas são originárias,enquanto objeto de pesquisa, de campos como a Estatística, Matemática Aplicada, Computação e Engenharia, sendo, muitas vezes, de difícil acesso ao campo das Ciências Sociais Aplicadas de maneira geral.

Dessa forma, este artigo possui doisobjetivos principais. O primeiro é de cunho didático, procurando contribuir com a discussão na área da cientometria no Brasil a respeito de técnicas analíticas que são ainda pouco utilizados por estudos na área, apresentando um estudo de caso concreto e sua forma de operacionalização. Esse objetivo visa explorar o potencial de uso de um método, no caso, a análise multivariada de componentes principais,que permita trabalhar múltiplas variáveis e entender como cada uma influencia o fenômeno analisado. O segundo é de interesse aplicado na questão da internacionalização da ciência brasileira, procurando produzir uma maneira de ler como as diferentes unidades de um sistema científico, no caso as universidades federais brasileiras, estão posicionadas em uma mesma base de dados. Para isso, utiliza-se como fonte de informação uma importante e relevante base de dados de publicações científicas, a *Web of Science*, de onde se extrairá a produção científica das59 universidades federais.

2 Análise multivariada de componentes principais

O interesse de aplicação e desenvolvimento de técnicas para análise multivariada tem apresentado um importante crescimento nas últimas décadas devido a duas razões principais, a insuficiência dos métodos tradicionais e a evolução da tecnologia da informação. Os pesquisadores de vários campos do conhecimento têm percebido a complexidade do comportamento humano e a insuficiência de suas técnicas tradicionais de pesquisa para os problemas que se propõem a resolver. A evolução da tecnologia da informação tem tornado mais fácil coletar uma vasta quantidade de informação, bem como mensurar novas variáveis a respeito de diferentes fenômenos humanos antes impossíveis (LATTIN; CARROLL; GREEN, 2011, p. 3).

Dessa maneira, muitas técnicas hoje podem ser classificadas como pertencentes à categoria de uma análise multivariada. As diferentes técnicas podem ser descritas a partir de três características fundamentais: a técnica é utilizada para a análise de interdependência entre as variáveis e os casos entre si ou para a dependência entre eles, a técnica é usada com objetivo de exploração dos dados ou com objetivo de confirmação e teste de uma hipótese e, por último, a técnica é projetada para lidar com dados métricos ou com dados não métricos (LATTIN; CARROLL; GREEN, 2011, p. 7). A partir dessas três características, organizam-se técnicas como componentes principais, análise fatorial exploratória, análise fatorial confirmatória, escalonamento multidimensional, análise de conglomerados, correlação canônica, modelos de equação estrutural com variáveis latentes, análise de variância, análise discriminante e modelos de escolha Logit. Neste artigo, damos ênfase apenas à análise multivariada de componentes principais.

Latin, Carroll e Green (2011, p.7) definem a análise de componentes principais da seguinte maneira:

> A análise de componentes principais é um método que pode ser usado para reduzir a dimensionalidade dos dados multivariados. Ela permite que o pesquisador reexpresse os dados (fazendo combinações lineares das variáveis originais) para que as primeiras poucas variáveis novas resultantes (chamadas componentes) respondam por tantas informações disponíveis quanto possível. Se uma redundância substancial estiver presente no conjunto de dados, então é possível explicar a maioria das informações do conjunto original de dados com um número relativamente pequeno de componentes. Essa redução da dimensão torna mais direta a visualização dos dados e sua subsequente análise, mais administrável.

A análise de componentes principais se aplica a conjuntos 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 lida com esse tipo de relacionamento apenas 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 à conclusão de 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 nos 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. É por estudar essas dimensões variáveis na caracterização do conjunto que propomos a utilização desta técnica para este artigo de cunho exploratório.

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 relativos a diferentes conjuntos de variáveis utilizados para descrever esses objetos. Para a realização de estudos comparativos tanto de indivíduos quanto 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 três etapas, sendo elas a definição e extração de dados, o tratamento e enriquecimento dos dados e, por fim, a 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 Educação[1] 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*, de modo a recuperar 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 três 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. Vale ressaltar aqui a importância do registro do campo "endereço" de origem de um pesquisador, sobretudo vinculando sua origem a sua instituição de trabalho. Somente a partir desse relacionamento torna-se possível recuperar a vinculação institucional dos artigos, mesmo que ele seja pertencente a várias instituições devido a colaborações de coautoria. Essa vinculação institucional a partir do campo "endereço" é reconhecida pela *Web of Science* e permite que o sistema faça um agrupamento de dados de forma automática, facilitando enormemente a coleta de indicadores já sistematizados e calculados a partir de uma mesma lógica de agregação por instituição.

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 sua identificação nos gráficos gerados e expostos a seguir. Vale ressaltar que esse tratamento teve um objetivo apenas estético, para reduzir ruído de informação desnecessária e sobreposição de texto nos gráficos apresentados a seguir. De forma a normalizar o uso do recurso de autocitação pelas universidades, também calculamos uma

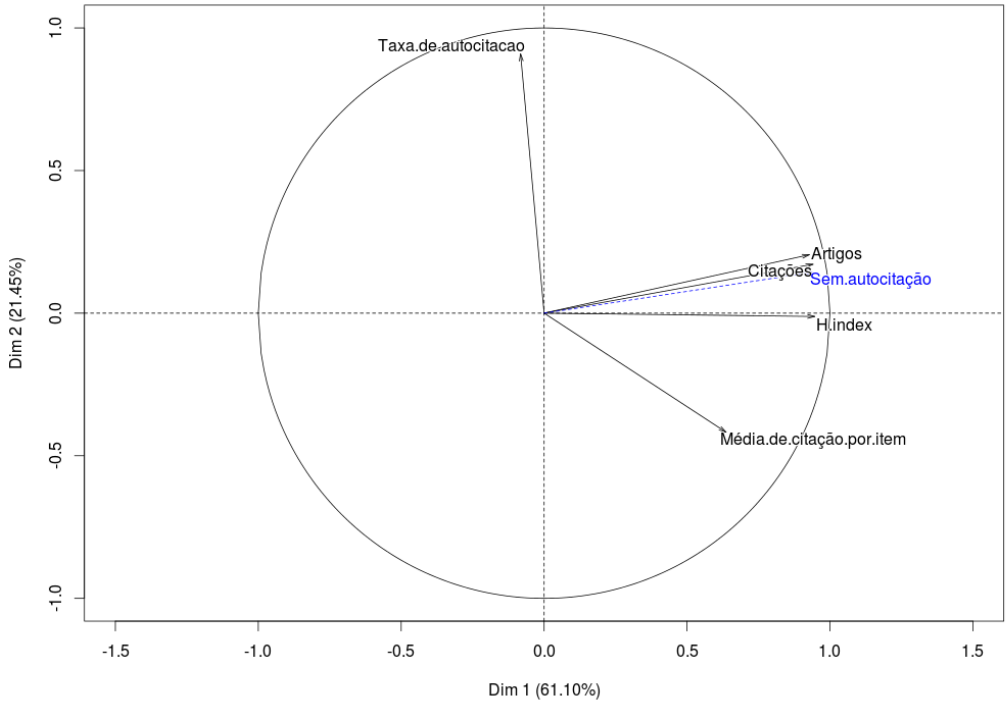
nova variável, chamada de taxa de autocitação, calculada pela razão entre o número de autocitações realizadas pelos pesquisadores de uma universidade e o número total de citações recebidas pelos seus artigos. Desse modo, incluímos uma nova variável a ser utilizada pela análise multivariada. O objetivo de incluir essa nova variável é perceber que efeito ela causa na relação com as outras variáveis deste estudo, procurando identificar se ela contribui para explicar possíveis diferenças entre as universidades federais e se a autocitação pode ser entendida como uma característica que é mais praticada por algumas universidades e menos por outras, revelando uma possível estratégia de articulação social em torno da produção científica dessas universidades.

Na etapa de realização da análise, utilizamos o *software* R^2 de estatística com o pacote FactoMineR, desenvolvido especialmente para a realização de diversos tipos de análise multivariada. O *software* foi escolhido por ser licenciado em formato *software* livre, o que facilita aos pesquisadores, estudantes e interessados na aplicação desse tipo de técnica aapropriação de ferramentas de apoio ao desenvolvimento de suas próprias pesquisas.

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 isso, temos que a distribuição e a estrutura de relacionamento das variáveis nos mostram três tendências relevantes a serem consideradas: as quatro variáveis, artigos, índice-H, número de citações e número de citações sem autocitações, são convergentes, ou seja, apresentam a mesma tendência entre elas; já a taxa de autocitação é praticamente ortogonal a essas quatro variáveis, demonstrando 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.

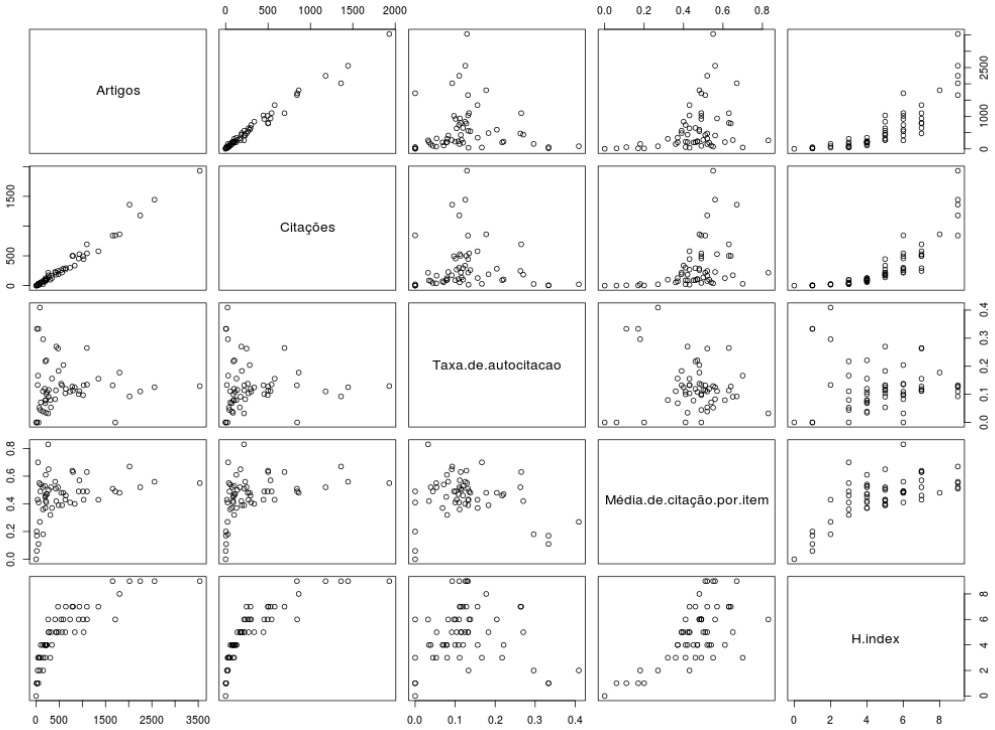
Figura 1 - Mapa de variáveis nas duas principais dimensões de análise



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 das correlações entre elas. Há correlações muito expressivas, como a Figura 1 já apresentou, como número de artigos 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.

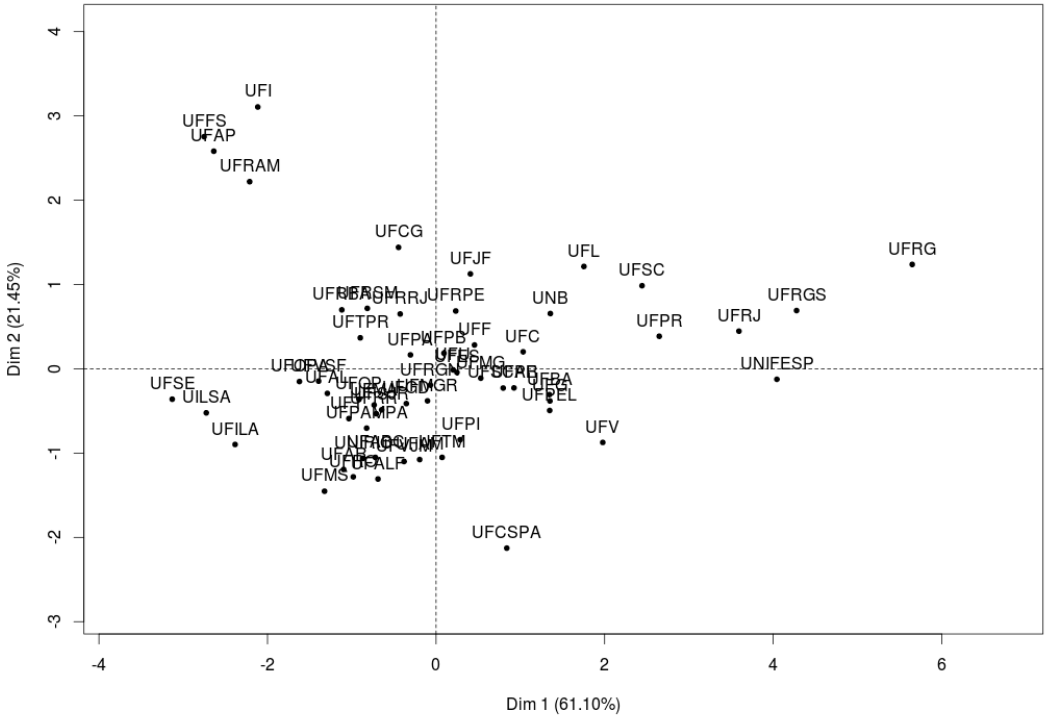
Figura 2 - Mapa de correlação entre as variáveis utilizadas



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 cinco 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 às 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 este estudo e que tipos de dados seriam redundantes em relação aos já recolhidos.

A criação de um novo indicador, a taxa de autocitação, se mostrou interessante como elemento que facilitou encontrar uma característica de variabilidade nos dados e, portanto, permitiu agrupar algumas instituições que se definiam de forma significativa por essa característica. Vale ressaltar que esse exercício de produzir novos indicadores e perceber seu

efeito junto às outras variáveis que compõem o estudo é uma importante característica a ser explorada a partir de técnicas de análise multivariada.

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, nota Capes, número de professores, entre outros.

Recomendamos a partir deste 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, dez. 2007.

HUSSON, François; LÊ, Sébastien; PAGÉS, Jérôme. **Exploratory multivariate analysis by example using R.**Boca Raton: CRC Press, 2011. 228p.

LATTIN, James; CARROLL, J. Douglas; GREEN, Paul E. **Análise de dados multivariados.**São Paulo: Cengage Learning, 2011. 455p.

LEYDESDORFF, Loet. **The challenge of Scientometrics**: the development, measurement, and self-organization of scientific communications. Leiden: Universal Publishers, 2001. 344p.

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.

em questão
E-ISSN 1808-5245
Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS
v. 20, n.3 – Edição Especial 2014

Using multivariate analysis for mapping the profile of internationalization of Brazilian federal universities: an exploratory study based on data available in the Web of Science database

Abstract: The paper aims to study the use of multivariate analysis in order to contribute to the discussion and dissemination of this type of technique in scientometric studies in Brazil. It also aims to apply this technique and analyze its possible application to study the internationalization of Brazilian scientific production focusing on the 59 universities in the federal system. The study uses data from the database Web of Science, builds a new indicator to assess the rate of self-citation of universities and conducts a multivariate analysis using the statistical computing free software R. The results allow us to propose the arrangement of the universities in five groups according to their features revealed by the variables analyzed. The method proves to be easy to use and very useful for scientometric studies.

Keywords: Multivariate analysis. Main components. Internationalization. Federal universities. Web of Science.

[1] Disponível em: http://portal.mec.gov.br/index.php.
[2] Disponível em: http://www.r-project.org.

Recebido: 27/07/2014
Aceito: 24/11/2014