UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Paula Lunardi de Mello

SISTEMÁTICAS DE AGRUPAMENTO DE PAÍSES COM BASE EM INDICADORES DE DESEMPENHO

Porto Alegre
2017

Paula Lunardi de Mello

Sistemáticas de agrupamento de países com base em indicadores de desempenho

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Produção.

Orientador: Michel José Anzanello, *Ph.D.*

Porto Alegre

2017

Paula Lunardi de Mello

Sistemáticas de agrupamento de países com base em indicadores de desempenho

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel José Anzanello, *Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. Flávio Sanson Fogliatto, *Ph.D*.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Danilo Marcondes Filho, Dr. (Estatística/UFRGS)

Professora Liane Werner, Dra. (PPGEP/UFRGS)

Professor Ricardo Augusto Cassel, Dr. (PMPEP/UFRGS)

DE MELLO, Paula Lunardi. ***Sistemáticas de agrupamento de países com base em indicadores de desempenho***, 2017. Dissertação (Mestrado em Engenharia) – Universidade Ferderal do Rio Grande do Sul, Brasil.

RESUMO

A economia mundial passou por grandes transformações no último século, as quais incluiram períodos de crescimento sustentado seguidos por outros de estagnação, governos alternando estratégias de liberalização de mercado com políticas de protecionismo comercial e instabilidade nos mercados, dentre outros. Figurando como auxiliar na compreensão de problemas econômicos e sociais de forma sistêmica, a análise de indicadores de desempenho é capaz de gerar informações relevantes a respeito de padrões de comportamento e tendências, além de orientar políticas e estratégias para incremento de resultados econômicos e sociais. Indicadores que descrevem as principais dimensões econômicas de um país podem ser utilizados como norteadores na elaboração e monitoramento de políticas de desenvolvimento e crescimento desses países. Neste sentido, esta dissertação utiliza dados do Banco Mundial para aplicar e avaliar sistemáticas de agrupamento de países com características similares em termos dos indicadores que os descrevem. Para tanto, integra técnicas de clusterização (hierárquicas e não-hierárquicas), seleção de variáveis (por meio da técnica "*leave one variable out at a time*") e redução dimensional (através da Análise de Componentes Principais) com vistas à formação de agrupamentos consistentes de países. A qualidade dos *clusters* gerados é avaliada pelos índices *Silhouette*, Calinski-Harabasz e Davies-Bouldin. Os resultados se mostraram satisfatórios quanto à representatividade dos indicadores destacados e qualidade da clusterização gerada.

Palavras-chave: Clusterização, Seleção de variáveis, Análise de Componentes Principais, Índices de Avaliação da Qualidade de Clusterização.

DE MELLO, Paula Lunardi. ***Countries clustering systematics based on performance indexes***, 2017. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

ABSTRACT

The world economy faced transformations in the last century. Periods of sustained growth followed by others of stagnation, governments alternating strategies of market liberalization with policies of commercial protectionism, and instability in markets, among others. As an aid to understand economic and social problems in a systemic way, the analysis of performance indicators generates relevant information about patterns, behavior and trends, as well as guiding policies and strategies to increase results in economy and social issues. Indicators describing main economic dimensions of a country can be used guiding principles in the development and monitoring of development and growth policies of these countries. In this way, this dissertation uses data from World Bank to elaborate a system of grouping countries with similar characteristics in terms of the indicators that describe them. To do so, it integrates clustering techniques (hierarchical and non-hierarchical), selection of variables (through the "leave one variable out at a time" technique) and dimensional reduction (appling Principal Component Analysis). The generated clusters quality is evaluated by the Silhouette Index, Calinski-Harabasz and Davies-Bouldin indexes. The results were satisfactory regarding the representativity of the highlighted indicators and the generated a good clustering quality.

Keywords: Clustering, Variable selection, Principal Component Analysis, Clustering Validation Measures.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ACP	Análise de Componentes Principais
BSA	*Business Software Alliance*
CH	Índice Calinski-Harabasz
DB	Índice Davies-Bouldin
GFCC	*Global Federation of Competitiveness Councils*
IDE	Índice de Desempenho Econômico
PIB	Produto Interno Bruto
PPC	Paridade do Poder de Compra
RNB	Rendimento Nacional Bruto
SI	*Silhouette Index*

SUMÁRIO

1 Introdução

1.1 Considerações Iniciais

O último século foi marcado por profundas mudanças na economia mundial. Avanços tecnológicos, alterações nos padrões monetários, crises financeiras, períodos de crescimento alternados com períodos de recessão e governos adotando ora estratégias liberais, ora protecionistas, tornaram o sistema econômico mais dinâmico, complexo e interdependente (ALMEIDA, 2001; SPERO & HART, 2010). Dificuldades no enfrentamento de tais cenários são esperadas, como aponta o relatório de Perspectivas Econômicas Globais do Banco Mundial, lançado em Junho de 2015. Pelo levantamento, o crescimento global não deverá ultrapassar a barreira dos 5,5% até 2017.

A análise de indicadores econômicos de desempenho de países se mostra uma fonte de informações importante, visto que tais indicadores são capazes de descrever padrões de comportamento e tendências de tais países, orientando governos e entidades competentes na geração de políticas e estratégias para incrementar resultados positivos (ou ainda evitar resultados negativos). Com a definição apropriada do perfil de determinados países (e eventuais agrupamentos dos mesmos), oferece-se subsídios para a gerência eficiente dos recursos disponíveis para investimento, adoção de políticas apropriadas e estabelecimento de acordos estratégicos com potenciais parceiros. Taylor e Schroeder (2015) acreditam na possibilidade de se desenvolver um sistema sensível a mudanças nos padrões de indicadores econômicos, a ponto de gerar um alerta precoce de ação. Economistas estimam que, com o devido tratamento às informações econômicas disponíveis, ganhos de eficiência poderiam ser gerados na ordem de US$ 15 trilhões ao PIB global até 2030 (*BSA Report*, 2015).

A Engenharia de Produção há tempos dedica estudos que buscam auxiliar as organizações a serem mais eficientes, sejam elas empresariais ou governamentais. Os indicadores de desempenho são importantes para a avaliação de instituições, pois são eles que mostram a evolução dos resultados para então ser possível melhorá-los. A partir desse diagnóstico, novas práticas organizacionais e estratégias podem ser tomadas para impactar os resultados positivamente.

No entanto, percebe-se que históricos de indicadores econômicos de desempenho de países são tipicamente volumosos e guardam informações não explícitas, mas extremamente

relevantes (TAYLOR & SCHROEDER, 2015). A aplicação de técnicas de análise multivariada de dados pode auxiliar na identificação de padrões nesses dados, permitindo a geração de ações apropriadas no âmbito de governo e de gestão.

Nesse sentido, esta dissertação utiliza dados disponibilizados pelo Banco Mundial para aplicar ferramentas de análise multivariada de dados na busca de agrupamentos de países com características similares quando avaliados a partir de indicadores. É composta por três artigos abordando clusterização, seleção de variáveis com propósito de clusterização e avaliação de indicadores de desempenho de agrupamentos de países.

1.2 Objetivos

O objetivo principal da dissertação é desenvolver sistemáticas de clusterização com vistas ao agrupamento de países a partir de indicadores econômicos.

Os objetivos específicos incluem:

- Selecionar os indicadores mais relevantes para clusterização de países utilizando a técnica "*leave one variable out at a time*";
- Aplicar a técnica de redução dimensional Análise de Componentes Principais (ACP) e avaliar a consistência dos agrupamentos gerados;
- Avaliar a qualidade de *clusters* gerados por meio da aplicação dos índices internos *Silhouette Index* (SI), Calinski-Harabasz (CH) e Davies-Bouldin (DB).

1.3 Justificativa do Tema

A gestão de desempenho organizacional é tema recorrente na Engenharia de Produção. Foi muito debatida e estudada desde que Kaplan e Norton (1997) apresentaram o *balanced scorecard,* que avalia o desempenho das organizações reunindo, além de indicadores financeiros, indicadores relativos a qualidade, produção, processos e produtos, entre outros. Os benefícios de uma análise mais abrangente e focada em resultados a longo prazo se mostraram consistentes e de ganhos expressivos.

Na administração pública não é diferente. Souza (2003) evidencia que no início da década de 90 uma nova agenda foi introduzida para os governos de muitos países, especialmente para os em desenvolvimento, e eles vêm tentando se adaptar às novas condições da economia mundial. Um plano estratégico bem definido pode guiar o caminho

para o desenvolvimento, mas somente o planejamento não garante o sucesso, é necessário que as estratégias sejam monitoradas e avaliadas sistematicamente (SOUZA, 2003).

Dados de indicadores são constantemente coletados. Com a evolução dos recursos computacionais, a velocidade e o volume com os quais os dados são criados têm crescido sistematicamente. Para Einav & Levin (2014), nas próximas décadas, o tratamento e processamento de grandes bases de dados mudará o panorama da política econômica e da pesquisa em economia. E dessa maneira a análise multivariada de dados tem muito a contribuir, convertendo variáveis em informações para as entidades tomadoras de decisão.

Diversos estudos têm utilizado a análise multivariada, principalmente áreas como medicina, psiquiatria, sociologia, antropologia, geologia, biologia, geografia, pesquisas de mercado e engenharia.

Quando se trata especificamente de indicadores econômicos, Camacho et al. (2008), Battisti e Parmeter (2013), De La Paz-Marín et al. (2015) são exemplos de estudos recentes na área, os quais avaliaram fenômenos de mercado e sua correlação com os indicadores econômicos. Tais estudos propuseram abordagens com vistas a extrair informações para criar estratégias de monitoramento que minimizem os efeitos nocivos à economia.

No âmbito teórico, Duarte et al. (2010) afirmam que pouco tem sido feito para validar partições de dados produzidas por métodos de clusterização. As técnicas de agrupamento têm sido amplamente estudadas, porém ainda há espaço para a avaliação dos agrupamentos originados. Métricas de avaliação de qualidade e comparação de resultados de diferentes índices podem trazer informações relevantes na validação de resultados, gerando soluções robustas e consistentes. De tal forma, percebe-se que o estudo aqui proposto justifica-se tanto no aspecto prático quanto teórico.

1.4 Procedimentos Metodológicos

Essa dissertação pode ser classificada como pesquisa exploratória, por buscar proporcionar maior familiaridade com o problema (GIL, 2002). Quanto à natureza, é pesquisa aplicada, pois objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos (DA SILVA, 2005). O estudo é uma pesquisa quantitativa devido à aplicação de análises numéricas.

No primeiro artigo foram estudados os dados de 1965 a 2014 de quatro indicadores econômicos para 205 países. As séries foram divididas em cinco décadas para possibilitar a observação de mudanças e tendências dos indicadores ao longo do tempo. Foi calculada a mediana de cada década para cada indicador e gerados os dendogramas. Na construção dos dendogramas foi utilizada a distância Euclidiana e o método de ligação de Ward, já aplicados em pesquisas anteriores relacionadas a indicadores econômicos (KUBRUSLY, 2001; MONTANARI et al., 2014). Analisando visualmente os dendogramas, foi definido o intervalo do número de *clusters* a ser testado para cada indicador nas diferentes décadas. O valor máximo do SI no intervalo foi usado para identificar o número ideal de *clusters* e assim gerar a clusterização pelo método *k-means*.

No segundo artigo foram usados dados de 1995 a 2014 de dez indicadores econômicos para 84 países, sobre os quais foram aplicados os seguintes passos: *i*) cálculo da mediana da série temporal dos indicadores por país; *ii*) construção e análise do dendograma dos dados, pelo método de ligação Ward; *iii*) geração de *clusters* pelo método *k-means* e remoção das variáveis menos relevantes para clusterização pelo critério do máximo SI médio; *iv*) eliminação do indicador não discriminante, removendo o indicador que conduziu ao máximo SI, de acordo com o procedimento "*leave one variable out at a time*". Repetição do passo (*iii*) até que um único indicador ficasse retido na base de dados.

O terceiro artigo utilizou a mesma base de dados do segundo, desta vez realizando análise de componentes principais com vistas à redução do número de indicadores. As componentes que explicavam pelo menos 90% da variância dos dados foram retidas na clusterização *k-means*. O intervalo de 2 a 42 *clusters* foi estudado e a consistência da solução foi avaliada através dos índices SI, Calinski-Harabasz e Davies-Bouldin. Ao final foi identificado o número recomendado de *clusters* buscando otimizar cada um dos índices.

1.5 Estrutura da Dissertação

A dissertação está estruturada em cinco capítulos. O primeiro capítulo traz a introdução do trabalho, além de apresentar os objetivos e as justificativas, bem como o método de pesquisa adotado. As delimitações e estrutura do trabalho finalizam o capítulo.

O segundo capítulo apresenta o primeiro artigo, que aplica técnicas hierárquicas e não hierárquicas aos indicadores econômicos dos países, formando *clusters* de países com perfis semelhantes em termos dos indicadores avaliados.

O terceiro capítulo traz o segundo artigo, com o objetivo de destacar os indicadores econômicos mais relevantes na formação de agrupamentos de países com desenvolvimento econômico semelhante. Integrando técnica de clusterização do tipo *k-means* ao procedimento "*leave one variable out at a time*" de seleção de variáveis, além de avaliar a qualidade dos *clusters* gerados através do SI..

O quarto capítulo apresenta o terceiro artigo, que é uma aplicação da ACP em indicadores econômicos de países, gerando na sequência agrupamentos através da técnica *k-means*. A qualidade dos agrupamentos gerados é avaliada através de três índices internos de avaliação de qualidade, onde a clusterização resultante da otimização de cada um é analisada.

Por fim, o quinto capítulo conclui o estudo, onde são avaliados os principais resultados frente aos objetivos traçados e as delimitações estabelecidas. Essa seção ainda traz sugestões para trabalhos futuros.

1.6 Delimitações do Estudo

São delimitações do presente estudo:

- As análises dos agrupamentos foram estritamente quantitativas, sendo excluídos pareceres sociais e políticos dos países (cunho qualitativo);
- Não houve desenvolvimento ou elaboração de novas técnicas e algoritmos para clusterização e seleção de variáveis, valendo-se de ferramentas existentes e fazendo sua integração para obter o resultado esperado;
- As variáveis são selecionadas com objetivo de clusterização, e não de predição ou classificação; e
- O banco de dados estudado se restringe aos indicadores econômicos coletados pelo Banco Mundial.

1.7 Referências

ALMEIDA, P. R. *A economia internacional no século XX: um ensaio de síntese.* Revista Brasileira de Política Internacional, vol.44, no.1, Jan./Jun, 2001.

BATTISTI, M.; PARMETER, C. F. *Clustering and polarization in the distribution of output: A multivariate perspective*. Journal of Macroeconomics, v. 35, p. 144-162, 2013.

CAMACHO, M.; PEREZ-QUIROS, G.; SAIZ, L. *Do European business cycles look like one?*. Journal of Economic Dynamics and Control, v. 32, n. 7, p. 2165-2190, 2008.

COSTA, M. I. *A avaliação de desempenho como motor de mudança e eficácia organizacional*. 2010. Tese de Doutorado.

DA SILVA, E. L.; MENEZES, E. M. *Metodologia da pesquisa e elaboração de dissertação*. UFSC, Florianópolis, 4a. edição, v. 123, 2005.

DE LA PAZ-MARÍN, M.; GUTIÉRREZ, P. A.; HERVÁS-MARTÍNEZ, C. *Classification of countries' progress toward a Knowledge Economy based on machine learning classification techniques*. Expert Systems with Applications, v. 42, n. 1, p. 562-572, 2015.

DUARTE, J. M. M. et al. *On consensus clustering validation*. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer Berlin Heidelberg, 2010. p. 385-394.

EINAV, L.; LEVIN, J. *The data revolution and economic analysis*. Innovation Policy and the Economy, v. 14, n. 1, p. 1-24, 2014.

GIL, A. C. *Como classificar as pesquisas*. _____. Como elaborar projetos de pesquisa, v. 4, p. 41-56, 2002.

KAPLAN, R. S.; NORTON, D. P. *A estratégia em ação: balanced scorecard*. 5. ed. Rio de Janeiro: Campos, 1997.

KUBRUSLY, L. S. *Um procedimento para calcular índices a partir de uma base de dados multivariados*. Pesqui. Oper., Rio de Janeiro , v. 21, n. 1, p. 107-117, jun. 2001

MONTANARI, M. G.; MIRANDA, R. A. M.; GIRALDI, J. M. E. *Uma comparação quantitativa entre dois índices de competitividade*. Revista de Globalización, Competitividad y Gobernabilidad, v. 8, n. 2, p. 121, 2014.

SOUZA, C. *Por que mudam as formas de gestão pública?* In: Gestão Pública: a trajetória da função administração do Estado da Bahia. Cadernos da Fundação Luís Eduardo Magalhães. Salvador: FLEM, 2003.

SPERO, J. E., HART, J. *The Politics of International Economic Relations*, Wadsworth (2010)

TAYLOR, L.; SCHROEDER, R. *Is bigger better? The emergence of big data as a tool for international development policy*. GeoJournal, v. 80, n. 4, p. 503-518, 2015

2 Primeiro Artigo: Análise de índices de desempenho de países através de ferramentas de clusterização

Paula Lunardi de Mello

Michel José Anzanello

Resumo

O último século foi de grandes transformações, principalmente na economia. Marcado por alternância entre períodos de crescimento e outros de estagnação, o mesmo acontecendo nas estratégias governamentais para enfrentá-los: ora liberalização do mercado, ora protecionismo comercial.. Visando enfrentar essas mudanças de cenário, a análise de indicadores de desempenho permite orientar políticas e estratégias para incrementar resultados positivos ou ainda evitar resultados negativos, sendo capazes de gerar informações relevantes a respeito de padrões de comportamento e tendências. Nesse sentido este artigo aplica ferramentas de clusterização nos dados do *Global Federation of Competitiveness Councils* (GFCC), as quais incluem técnicas hierárquicas e não hierárquicas, com visitas à formação de agrupamentos de países com perfis semelhantes em termos dos indicadores avaliados. A clusterização criada a partir de valores de SI se mostrou adequada para indicadores econômicos, mostrando que apesar do contexto econômico e social de cada país não fazer parte da análise, existe distinção entre os *clusters* formados

Palavras-chave: indicadores econômicos, clusterização, *Silhouette Index*.

2.1 Introdução

A economia mundial passou por grandes transformações no último século. Almeida (2001) cita os avanços tecnológicos, as mudanças de padrões monetários, as crises financeiras, os períodos de crescimento sustentado seguidos de outros de estagnação, governos alternando estratégias de liberalização do mercado com políticas de protecionismo comercial. No final do século XX, essas profundas mudanças, principalmente no âmbito

político, derrubaram as barreiras geográficas entre países e a competitividade passou a ser global, tornando o sistema econômico mais dinâmico, complexo e interdependente (SPERO & HART, 2010).

As projeções têm indicado tempos difíceis para os próximos anos. Segundo o relatório de Perspectivas Econômicas Globais do Banco Mundial, lançado em Junho de 2015, o crescimento global não deverá ultrapassar a barreira dos 5,5% até 2017. Diretrizes e soluções para enfrentar períodos desfavoráveis à atividade econômica devem ser desenvolvidas pelos governos. Uma alternativa é a análise de indicadores de desempenho, capazes de gerar informações relevantes a respeito de padrões de comportamento e tendências, permitindo orientar políticas e estratégias para incrementar resultados positivos ou ainda evitar resultados negativos.

A não identificação do perfil de um determinado país em termos de seus indicadores de desempenho pode levar a diversos problemas, tais como: i) gestão ineficiente dos recursos disponíveis para investimentos, uma vez que setores de maior impacto no desempenho geral não são priorizados; ii) adoção de políticas equivocadas, pela falta de conhecimento de como é composto o desempenho geral; e iii) estabelecimento de acordos não estratégicos, pelo não reconhecimento de potenciais parceiros.

A análise de dados históricos através de ferramentas multivariadas se apresenta como uma alternativa eficiente para avaliação de indicadores. A partir dela é possível reconhecer padrões nas variáveis de estudo (no caso, os indicadores), objetivando classificar e agrupar os países de acordo com seu desempenho geral. Parte-se do princípio que a análise de agrupamentos é mais direta e eficiente do que a análise individual das observações isoladas. Alinhado com esse objetivo, Rencher (2002) cita a clusterização como uma abordagem adequada para tal fim, visto que permite identificar padrões dos países (observações) agrupados em *clusters*, de forma que a similaridade em observações de um mesmo grupo e a diferença entre grupos é determinada. Tal ferramenta tem sido aplicada em diversos segmentos da ciência, incluindo medicina, psiquiatria, sociologia, antropologia, geologia, geografia, pesquisas de mercado, engenharia e economia (RENCHER, 2002).

Alinhado com as proposições acima, este artigo aplica ferramentas de clusterização nos dados do Banco Mundial, as quais incluem técnicas hierárquicas e não hierárquicas, com

visitas à formação de agrupamentos de países com perfis semelhantes em termos dos indicadores avaliados. Objetiva-se monitorar países que apresentam valores extremos nos indicadores, os quais tendem a originar *clusters* isolados (constituídos por uma única observação), além de analisar a evolução do desempenho de determinados países ao logo do tempo.

O trabalho está dividido em cinco seções. A seção 2.2 apresenta a revisão da literatura no que diz respeito a ferramentas multivariadas, aprofundando a pesquisa em métodos de clusterização. A seção 2.3 apresenta o método aplicado para a análise dos dados de desempenho dos países. Já na seção 2.4 são mostrados os resultados obtidos pela aplicação do método aos dados. Por fim, na seção 2.5 são expostas as conclusões geradas pelo estudo.

2.2 Revisão da literatura

A revisão da literatura consiste na pesquisa por clusterização, análise de *cluster* e suas aplicações.

2.2.1 Análise de *cluster*

A análise multivariada consiste na análise simultânea de duas ou mais variáveis (JOHNSON & WICHERN, 1992; HAIR et al., 2005). Permite identificar padrões ou realizar testes de hipóteses de um determinado interesse, através da estatística descritiva e inferencial dos dados (RENCHER, 2002). Kubrusly (2001) demonstra que a análise multivariada de dados reúne as ferramentas adequadas para avaliação de índices como os analisados neste estudo e para a formação de agrupamentos.

Rencher (2002) destaca que há dois objetivos para a separação de grupos: descrever ou explicar as diferenças entre dois ou mais grupos e realizar a atribuição de observações a grupos.

Para Hair et al. (2005), os agrupamentos resultantes devem apresentar elevada homogeneidade dentro dos grupos e elevada heterogeneidade entre os grupos formados. Se a classificação for realizada de maneira correta, quando representada graficamente, as observações de um mesmo grupo estarão próximas e as de diferentes grupos estarão distantes (HAIR et al., 2005). Essa técnica é usada principalmente com objetivo exploratório, pois cria grupos sem observar qualquer estrutura pré-existente nos dados, e a solução de agrupamento

depende da base usada como medida de similaridade, ou seja, as variáveis (HAIR et al., 2005). Ainda segundo Hair et al. (2005), é grande a dependência dos grupos à base de dados estudada, tendo o acréscimo e a eliminação de variáveis um grande impacto sobre a solução resultante.

A clusterização é uma técnica que realiza o agrupamento de itens em função das similaridades ou distâncias (dissimilaridades) entre observações de uma variável (JOHNSON & WICHERN, 1992). Já para Cruz e Ochi (2011), clusterização é o processo que integra objetos semelhantes de uma base de dados em um mesmo grupo (mesmo *cluster*). Problemas onde o número de clusters é previamente conhecido são chamados de *k*-clusterização ou Problemas de Clusterização simplesmente; se ele for desconhecido, o problema é denominado Problema de Clusterização Automática (CRUZ & OCHI, 2011). Ambos os problemas podem ser classificados como NP-Hard, mas o fato de o número de clusters não ser previamente conhecido aumenta substancialmente o número de soluções possíveis e torna o problema ainda mais complexo (CRUZ & OCHI, 2011).

Hair et al. (2005) dividem as técnicas de agrupamentos em dois tipos de procedimentos: hierárquicos e não-hierárquicos. O primeiro tipo envolve a construção de uma hierarquia com estrutura do tipo árvore, que por sua vez tem também uma distinção entre métodos aglomerativos e divisivos (HAIR et al., 2005). Nos métodos aglomerativos, cada objeto ou observação começa como seu próprio grupo e, nos passos seguintes, os dois grupos mais próximos são combinados em um novo agregado, reduzindo o número de agrupamentos em uma unidade a cada passo (HAIR et al., 2005). À medida que os passos de aglomeração se sucedem, todas as observações podem ser reunidas em um grande agregado e, por essa razão, os procedimentos aglomerativos são chamados de métodos construtivos (HAIR et al., 2005).

A lógica contrária é empregada nos métodos divisivos, onde inicialmente todas as observações estão em um único aglomerado e, em passos sucessivos, as observações mais diferentes entre si se separam e são transformadas em outros grupos menores construtivos (HAIR et al., 2005). Johnson e Wichern (1992) esclarecem que o resultado de ambos os métodos, aglomerativo e divisivo, pode ser representado em um gráfico de duas dimensões, chamado de dendograma, sendo possível identificar as junções e divisões que foram realizadas nos diferentes níveis, como ilustra a Figura 2.1.

Hair et al. (2005) enumeram os algoritmos aglomerativos mais aplicados, os quais incluem o de ligação individual, de ligação completa, de ligação média, o método de Ward e o método centroide. Cada método difere na forma de calcular a distância entre os agrupamentos.

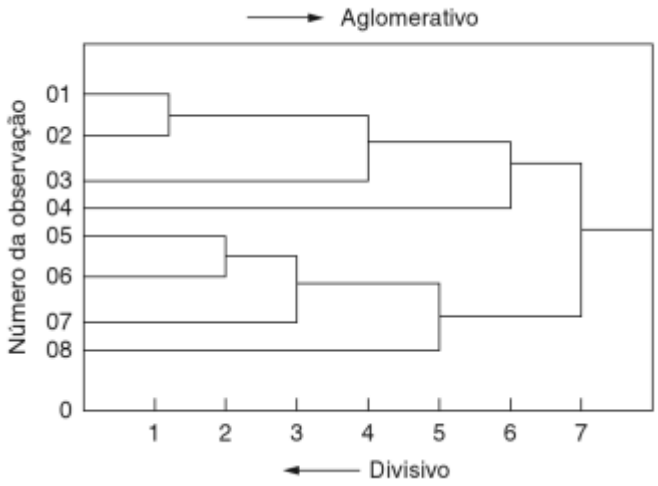

Figura 2.1: Dendograma ilustrando agrupamento hierárquico (HAIR et al., 2005)

Os procedimentos não-hierárquicos de agrupamento, por sua vez, não envolvem o processo de construção em árvore, pois partem de uma semente de agrupamento como centro inicial de um agregado, e todas as observações dentro de uma distância de referência pré-especificada são incluídas em um mesmo agrupamento (HAIR et al., 2005). De tal forma, cabe ao decisor conhecer previamente o número de clusters a ser gerado.

A crescente utilização dos métodos não-hierárquicos se justifica por serem menos suscetíveis a observações atípicas presentes nos dados. O método *k-means* é o mais comum deles (HAIR et al., 2005), o qual parte de um número inicial de *k* clusters e agrupa cada observação ao centroide (valor médio das observações contidas no agrupamento) mais próximo (JOHNSON & WICHERN, 1992). Este método tipicamente utiliza a distância Euclidiana para identificar diferentes clusters nos dados (LEE, 2013), embora outras possam ser utilizadas.

Como destacado por Rencher (2002), a distância euclidiana é uma medida muito utilizada para avaliar a similaridade entre observações. Considere duas observações (objetos a serem clusterizados) descritas pelos vetores $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_p)$, ambos

consistindo de *p* variáveis. A distância euclidiana entre eles é apresentada na Eq. (2.1), sendo que a soma das distâncias euclidianas entre elementos classificados no mesmo agrupamento deve ser minimizada em um procedimento de clusterização.

$$d\ (\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} = \sqrt{\sum_{j=1}^{p}(x_j - y_i)^2} \tag{2.1}$$

Uma técnica para validação do desempenho da clusterização é o S*ilhouette Index* (SI), que se baseia nas diferenças entre as distâncias das observações, analisadas aos pares dentro do *cluster* (LIU ET AL., 2010). Esse índice varia de -1 a +1 e o número ótimo de clusters é determinado pela maximização desse índice, que mede o quão distintos os *clusters* formados estão uns dos outros.

Para Anzanello e Fogliatto (2011), o SI_n é calculado de acordo com a Eq. (2.2):

$$\mathrm{SI}_n = \frac{b(n) - a(n)}{\max\{b(n), a(n)\}} \tag{2.2}$$

sendo *a(n)* a média das distâncias da *n*-ésima observação dentro do mesmo *cluster*, e *b(n)* a média das distâncias da mesma observação em relação as outras alocadas no *cluster* mais próximo.

O SI não depende da técnica de clusterização utilizada na análise, uma vez que é calculado com base em distâncias. Normalmente é usado para medir a qualidade global da clusterização através da média de SI_n, conforme Eq. (2.3):

$$\overline{\mathrm{SI}} = \frac{\sum_{n=1}^{N} \mathrm{SI}_n}{N} \tag{2.3}$$

onde *n* o índice da observação e *N* o total de observações.

Já a distância Euclidiana pode ser uma medida de similaridade adequada para análise de clusters em indicadores econômicos, por ter sido aplicada com bons resultados em outros estudos, como os de Kubrusly (2001), Camacho et al. (2008) e Apergis et al. (2012).

Ao mesmo tempo, Hennig e Liao (2013) ponderam que não há clusters únicos em um conjunto de dados, e que selecionar um método de agrupamento adequado exige adequar as características analíticas do agrupamento resultante com os requisitos dependentes do contexto. Hennig e Liao (2013) ainda afirmam que esperar que os dados apresentem sua verdadeira estrutura através da aplicação de métodos automáticos pode ser um engano.

Por outro lado, Montanari et al. (2014) comprovaram que o impacto de uma variável analítica, ou qualitativa, na formação dos clusters pode não ser significativa. Montanari et al. (2014) compararam diferentes índices de competitividade de países em caráter exploratório, realizando uma análise de *cluster* do tipo hierárquico antes de uma não-hierárquica, usando a distância Euclidiana como base. Outro estudo que evitou o uso de pressupostos qualitativos na análise de *clusters* foi o de Stevans et al. (2012), realizando análise de componentes principais para definir o peso das variáveis econômicas, e assim ranquear países em relação à sua competitividade.

2.2.2 Aplicações de análise de *cluster* em indicadores de desempenho

A clusterização é comumente aplicada na análise de indicadores econômicos e sociais de países. Um exemplo é o estudo de Apergis et al. (2012), que aplicaram a clusterização para realizar a formação de grupos de países. O objetivo da pesquisa era capturar os efeitos do desenvolvimento do setor financeiro sobre a renda per capita real em 50 diferentes países. Dados de mais de dez variáveis foram acessados da base de dados do Banco Mundial, e o modelo de fator de variação de tempo não-linear proposto por Phillips e Sul (2007) foi aplicado. Os resultados indicaram que as medidas de desenvolvimento financeiro podem explicar parte da renda real per capita.

Camacho et al. (2008) analisaram os ciclos econômicos na Europa para identificar se cada país europeu é semelhante o suficiente para validar o que alguns autores chamaram de “ciclo europeu”. Adotaram a análise de *cluster* calculando a distância euclidiana entre as características dos países e não encontraram semelhanças claras entre seus ciclos.

De La Paz-Marín et al. (2015) utilizaram um método hierárquico de clusterização para agrupar 54 países de acordo com seu estágio na economia do conhecimento. A economia do conhecimento, por sua vez, é tida como a produção de bens e serviços com base em atividades intensivas em conhecimento que contribuam para um ritmo acelerado de avanço técnico e científico, bem como rápida obsolescência (POWELL & SNELLMAN, 2004). No estudo de De La Paz-Marín et al. (2015) o objetivo foi oferecer uma ferramenta gerencial para apoiar a tomada de decisão, permitindo identificar e analisar a evolução de um país na economia do conhecimento.

Por sua vez, a pesquisa realizada por Kubrusly (2001) analisou oito variáveis socioeconômicas de países da América Latina, usando como técnica de agrupamento o método Ward e distância euclidiana. O índice obtido foi considerado como um indicador de desenvolvimento econômico dos diferentes *clusters*.

Por fim, Battisti e Parmeter (2013) analisaram a renda per capita de 74 países e compararam resultados da clusterização utilizando técnicas bivariada e multivariada. Reforçam que as conclusões de uma perspectiva multivariada são mais ricas e trazem uma visão mais completa do panorama objeto de estudo.

2.3 Método

O estudo foi realizado utilizando dados reunidos pelo GFCC. Essa federação é uma rede de líderes de conselhos de competitividade em nível mundial e define direcionadores (*drivers*) fundamentais da competitividade nacional que sustentam o crescimento econômico global. Esses *drivers* são: performance geral, complexidade econômica, infraestrutura, talento, capital, inovação, qualidade de vida e crescimento futuro. Para a análise deste artigo, foram escolhidos quatro indicadores econômicos agrupados pertencentes ao *driver* performance geral, apresentados no Quadro 2.1.

A seleção dos indicadores se deu com base na sua representatividade para caracterizar a economia dos países. No caso das exportações, elas geram renda e se passam como investimento doméstico (RESENDE & VIEIRA, 2011). O Produto Interno Bruto (PIB), indica o nível de riqueza do país (AUSLOOS & LAMBIOTTE, 2007). Por consequência, o crescimento do PIB e a inflação foram selecionados por terem relação com o último indicador; de acordo com López-Villavicencio e Mignon (2011), o crescimento do PIB é influenciado pelo nível de inflação, apesar da relação não ser linear.

Quadro 2.1: Indicadores da performance geral da competitividade dos países

Indicador	**Medida**	**Definição**
Exportação de bens e serviços	Percentual do Produto Interno Bruto	Exportações de bens e serviços representam o valor de todos os bens e outros serviços de mercado fornecidos ao resto do mundo. Eles incluem o valor das mercadorias, fretes, seguros, transportes, viagens, royalties, taxas de licença, e outros serviços, tais como comunicação, construção, financeiro, informação, negócio, pessoal, e os serviços governamentais. Exclui a compensação de funcionários e rendimentos de investimentos (anteriormente chamado de fator de serviços) e pagamentos de transferências.
Produto Interno Bruto (PIB) per capita	Moeda, em dólares	PIB per capita é o produto interno bruto dividido pela população na metade do ano. O PIB é a soma do valor agregado bruto por todos os produtores residentes na economia mais quaisquer impostos sobre os produtos e menos quaisquer subsídios não incluídos no valor dos produtos; é calculado sem fazer deduções para depreciação de ativos fabricados ou para o esgotamento e a degradação dos recursos naturais
Crescimento do PIB	Anual, em percentual	Taxa anual de crescimento percentual do PIB a preços de mercado com base em moeda local constante
Inflação dos preços ao consumidor	Anual, em percentual	Inflação medida pelo índice de preços ao consumidor que reflete a variação percentual anual no custo para o consumidor médio de aquisição de uma cesta de bens e serviços que possam ser fixados ou alterados em intervalos específicos, tais como anual

Fonte: Banco Mundial

Esta pesquisa pode ser classificada como de natureza aplicada, utilizando abordagem quantitativa, pois analisa dados reais da economia dos países em questão, focada em dados. Seu objetivo é exploratório, por buscar proporcionar maior familiaridade com o tema (GIL, 2002).

Os dados foram obtidos diretamente do sítio do Banco Mundial e abrangeram 205 países. A série temporal inicia em 1965 e termina em 2014, e foi dividida em cinco décadas para possibilitar a observação de mudanças e tendências dos indicadores ao longo do tempo.

A clusterização realizada parte da mediana dos dados de cada década, por ser uma estatística menos suscetível a valores discrepantes. Para a construção dos dendogramas de cada indicador, foi utilizada a distância euclidiana e o método de ligação de Ward, já aplicados em pesquisas anteriores relacionadas a indicadores econômicos (KUBRUSLY, 2001; MONTANARI et al., 2014). Todos os cálculos necessários para a clusterização dos dados e elaboração dos dendogramas foram realizados no software Matlab, amplamente utilizado para análise de dados quantitativos (SILVA, 2004).

Apesar do número máximo de *clusters* gerados pelo Matlab nos dendogramas, por padrão, ser definido como 30, foram gerados como teste dendogramas sem essa limitação. Porém, o resultado não foi satisfatório, uma vez que a distância entre os *clusters* diminuiu significativamente, dificultando a análise visual dos diagramas e por consequência a identificação do intervalo de *k* a ser estudo. Dessa forma, foi adotado o padrão do software para gerar os dendogramas.

Analisando os dendogramas, foi identificado o intervalo do número de *clusters* a ser testado para cada indicador nas diferentes décadas. A partir daí foi realizada uma avaliação do número ideal de *clusters* pelo máximo valor do SI para método *k-means*, através da função *evalclusters* do Matlab.

2.4 Resultados

Os dendogramas apresentam no eixo das abscissas os países e na ordenada as distâncias entre seus agrupamentos. Os dendogramas construídos a partir dos dados do indicador de Exportação para as diferentes décadas são mostrados nas Figuras 2.2 a 2.6.

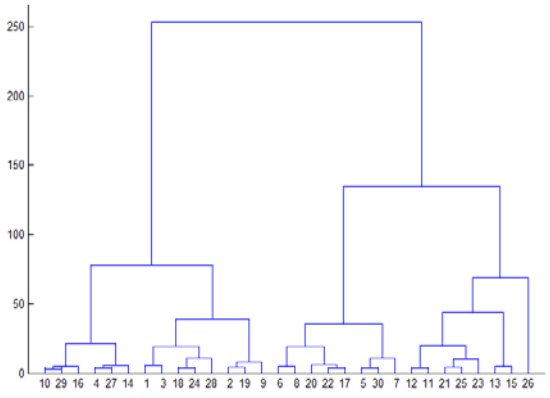

Figura 2.2: Dendograma para Exportação (1965-1974)

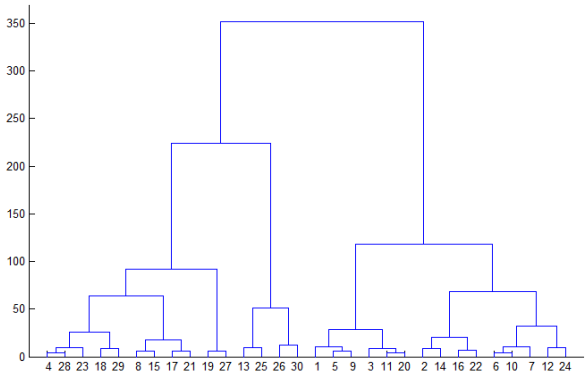

Figura 2.3: Dendograma para Exportação (1975-1984)

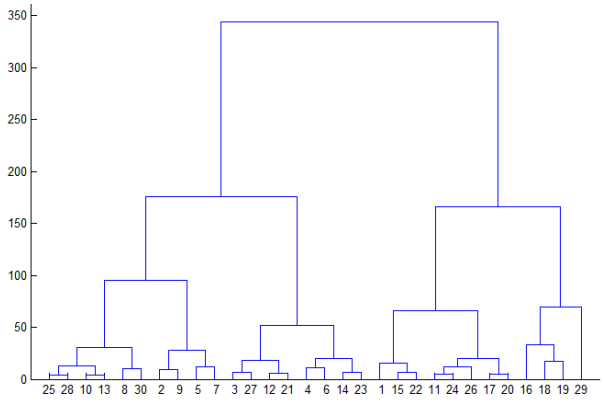

Figura 2.4: Dendograma para Exportação (1985-1994)

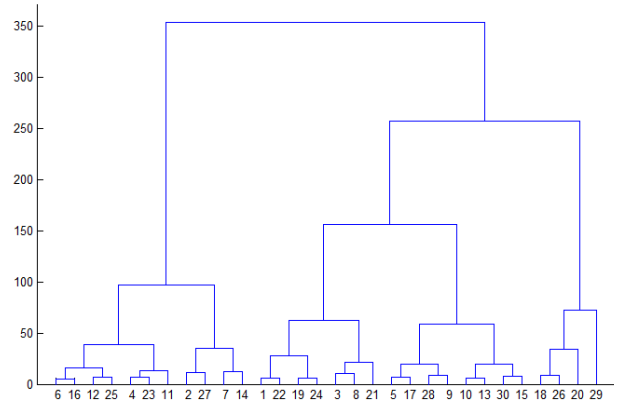

Figura 2.5: Dendograma para Exportação (1995-2004)

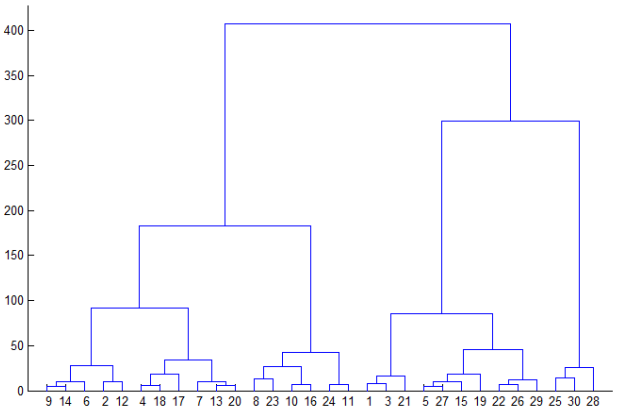

Figura 2.6: Dendograma para Exportação (2005-2014)

Nota-se pelas distâncias do eixo das ordenadas que basicamente o indicador *Exportação* apresenta quatro *clusters* distintos ao longo das décadas. A exceção é a década apresentada na Figura 2.3, onde parece existir um número maior de *clusters*. Portanto, será

testado o intervalo de 2 a 4 para as décadas das Figuras 2.2, 2.4, 2.5 e 2.6 e será usado o intervalo de 2 a 6 *clusters* para a década da Figura 2.3.

Na Tabela 2.1 são apresentados valores do SI para diferentes intervalos de números de *clusters*.

Tabela 2.1 - SI para *k-means* – Indicador Exportação

Década	**Núm. observações**	**Núm. *clusters* analisados**	**SI**	***k* recomendado**
1965-1974	115	[2 3 4]	[0,8074 0,7451 0,7143]	2
1975-1984	147	[2 3 4 5 6]	[0,7375 0,7779 0,6809 0,7153 0,6952]	3
1985-1994	183	[2 3 4]	[0,7194 0,7341 0,7066]	3
1995-2004	195	[2 3 4]	[0,7174 0,7527 0,7224]	3
2005-2014	194	[2 3 4]	[0.7378 0.7459 0.7478]	4

Pela Tabela 2.1, vê-se que a clusterização com a formação de 3 grupos conduziu aos melhores resultados em parte do período analisado. Já na última década é possível verificar que o arranjo recomendado foi de quatro *clusters*. A Tabela 2.2 traz os dados dos *clusters* formados com o número ótimo *k* encontrado pelo valor do SI de cada uma das décadas.

Tabela 2.2 - Dados dos *clusters* formados com *k* recomendado – Indicador Exportação (em % do PIB)

Indicador Exportação	**1965-1974**	**1975-1984**	**1985-1994**	**1995-2004**	**2005-2014**
Total de países	115	147	183	195	194
Qtde países *Cluster* 1	90	48	107	70	72
Qtde países *Cluster* 2	25	91	14	117	83
Qtde países *Cluster* 3	-	8	62	8	36
Qtde países *Cluster* 4	-	-	-		3
Valor máximo	123,310	166,200	165,085	182,850	214,094
Valor mínimo	3,328	2,764	1,938	0,439	5,771
Mediana	21,809	27,2619	28,060	33,205	38,363
País do vlr máx	Cingapura	Cingapura	Cingapura	Cingapura	Hong Kong
País do vlr mín	China	Laos	Polinésia Francesa	Mianmar	Eritreia
País mediana	Reino Unido	Tonga	Butão	Argélia	Líbano
Exemplo país *cluster* 1'	Síria	Dominica	Nicarágua	Moldávia	Fiji
Exemplo país *cluster* 2'	Barbados	El Salvador	Turquemenistão	Samoa	Turquia
Exemplo país *cluster* 3'	-	Panamá	São Cristóvão e Nevis	Cingapura	Emirados Árabes Unidos
Exemplo país *cluster* 4'	-	-	-	-	Luxemburgo

Na primeira década, os países com maiores taxas de exportação de bens e serviços figuram no *cluster* 2. Destaque para Cingapura que se manteve por quatro décadas como país com maior taxa de exportação de bens e serviços, sendo ultrapassado por Hong Kong apenas na última década. Outro destaque fica por conta do desempenho da China, que na década de 1965 a 1974 apresentava a menor taxa de exportação em relação ao PIB (3,32%) e no final do levantamento elevou essa taxa a 25,81%.

Os dendogramas gerados para o indicador *PIB per capita* são mostrados nas Figuras 2.7 a 2.11.

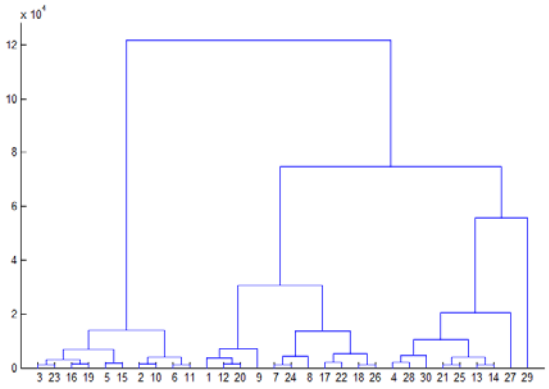

Figura 2.7: Dendograma para PIB per capita (1965-1974)

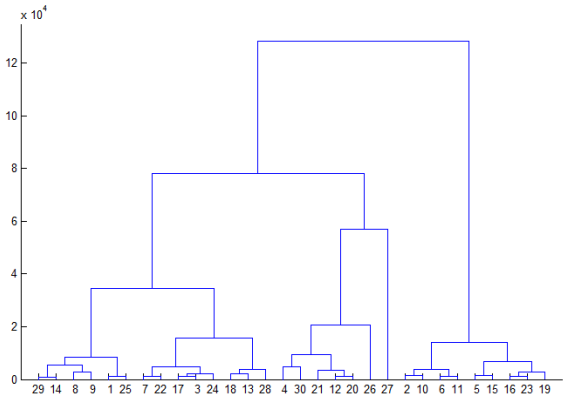

Figura 2.8: Dendograma para PIB per capita (1975-1984)

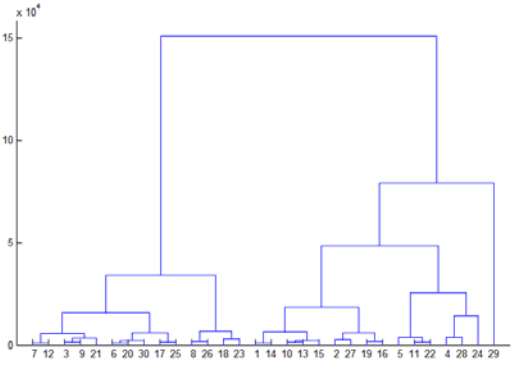

Figura 2.9: Dendograma para PIB per capita (1985-1994)

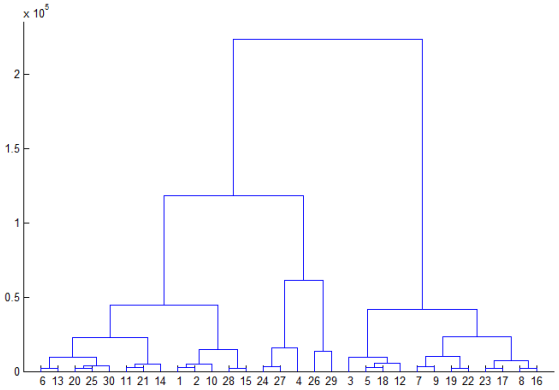

Figura 2.10: Dendograma para PIB per capita (1995-2004)

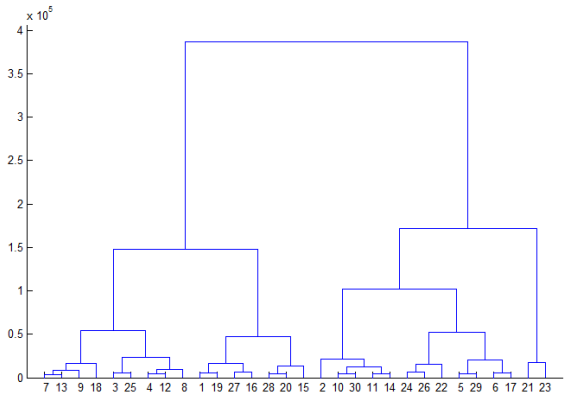

Figura 2.11: Dendograma para PIB per capita (2005-2014)

Nos dendogramas do indicador *PIB per capita* também se verifica que as maiores distâncias se dão em k = 4, mas na década de 1995 a 2004 vê se que a estrutura de agrupamento é diferente das demais décadas, apresentando maior número de grupos. Por isso para as décadas das Figuras 2.7 a 2.9 e 2.11 será testado intervalo de 2 a 4 *clusters*, e para Figura 2.10 será testado intervalo de 2 a 6.

A Tabela 2.3 apresenta o valor do SI para o indicador *PIB per capita* nas diferentes décadas.

Tabela 2.3 - SI para *k-means* – Indicador PIB per capita

Década	**Núm. observações**	**Núm. *clusters* analisados**	**SI**	***k* recomendado**
1965-1974	187	[2 3 4]	[0,8554 0,8414 0,8893]	4
1975-1984	188	[2 3 4 5]	[0,8647 0,8478 0,8877 0,8684]	4
1985-1994	191	[2 3 4 5]	[0,8479 0,8995 0,8881 0,8541]	3
1995-2004	203	[2 3 4 5 6]	[0,8515 0,9029 0,8675 0,7794 0,7855]	3
2005-2014	205	[2 3 4]	[0,8609 0,8944 0,8300]	3

Na Tabela 2.3 é possível verificar uma diminuição na variabilidade das observações ao final do intervalo analisado, indicando um número menor de *clusters* a serem gerados. A partir destes dados, são formados os *clusters* com base no número recomendado de *clusters* e as informações estão reunidas na Tabela 2.4.

Mônaco é o país que se mantém com maior PIB per capita por todo o intervalo observado, tendo um aumento significativo neste indicador. Por seu valor ser bastante distinto das demais observações, esse país se manteve sozinho em um *cluster*.

Tabela 2.4 - Dados dos *clusters* formados com *k* recomendado – Indicador PIB per capita (em USD)

Ind. PIB per capita	**1965-1974**	**1975-1984**	**1985-1994**	**1995-2004**	**2005-2014**
Total de países	187	188	191	203	205
Qtde países *Cluster* 1	1	29	2	42	38
Qtde países *Cluster* 2	12	1	37	4	162
Qtde países *Cluster* 3	31	145	152	157	5
Qtde países *Cluster* 4	143	13	-	-	-
Valor máximo	65738,434	68500,725	76150,647	92254,646	152877,365
Valor mínimo	149,004	149,722	150,952	124,364	202,312
Mediana	1515,104	1527,835	1486,858	2163,422	4692,869
País do vlr máx	Mônaco	Mônaco	Mônaco	Mônaco	Mônaco
País do vlr mín	Somália	Eritreia	Eritreia	Etiópia	Burundi
País mediana	Cazaquistão	São Vicente e Granadinas	Romênia	Peru	Argélia
Exemplo país *cluster* 1'	Mônaco	Nova Caledônia	Liechtenstein	Suécia	Dinamarca
Exemplo país *cluster* 2'	Suécia	Mônaco	Alemanha	Liechtenstein	Haiti
Exemplo país *cluster* 3'	Andorra	Botsuana	Chile	Belize	Noruega
Exemplo país *cluster* 4'	Belize	Dinamarca	-	-	-

Também na Tabela 2.4 se verifica que quatro países africanos se revezam com o valor mínimo. É possível observar que parcela significativa dos países se concentra em um único *cluster*, na sua maioria países em desenvolvimento.

Os dendogramas gerados para o indicador *Crescimento do PIB* são mostrados nas Figuras 2.12 a 2.16.

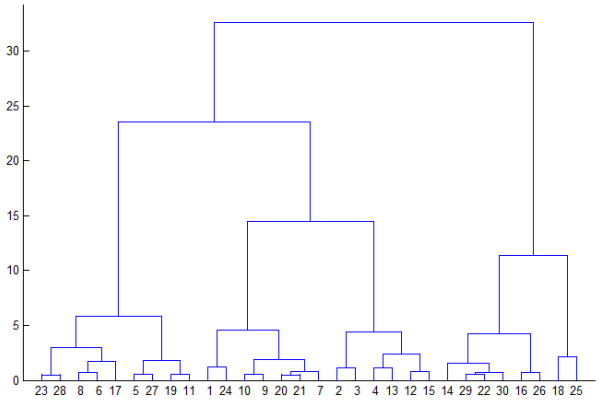

Figura 2.12: Dendograma para Crescimento do PIB (1965-1974)

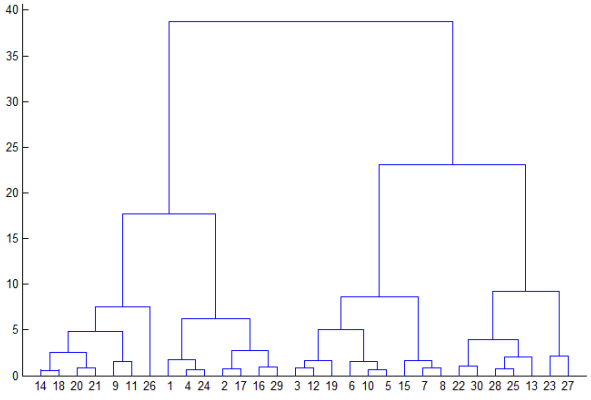

Figura 2.13: Dendograma para Crescimento do PIB (1975-1984)

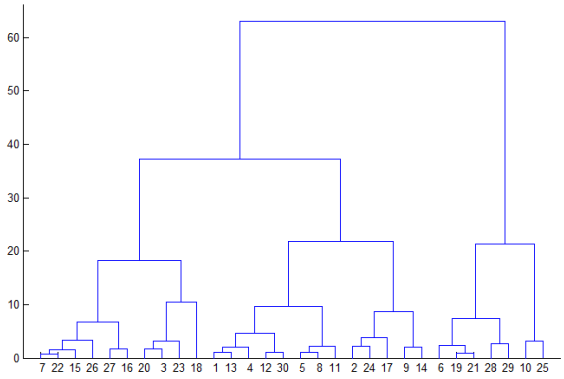

Figura 2.14: Dendograma para Crescimento do PIB (1985-1994)

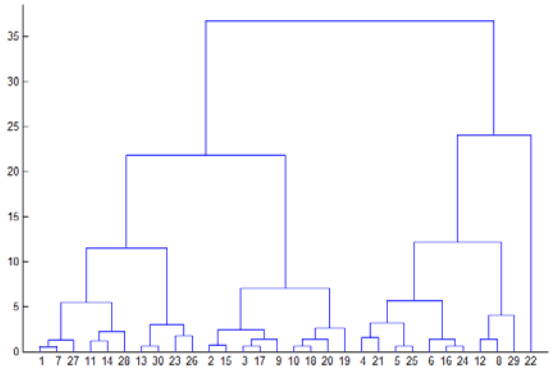

Figura 2.15: Dendograma para Crescimento do PIB (1995-2004)

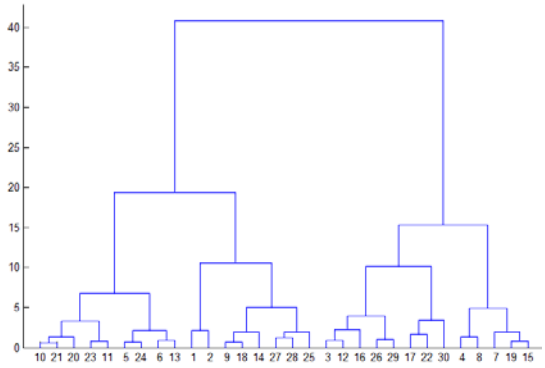

Figura 2.16: Dendograma para Crescimento do PIB (2004-2014)

Diferente dos indicadores anteriores, o *Crescimento do PIB* apresenta estrutura de agrupamento semelhante ao longo das décadas, sendo possível identificar até 6 *clusters* nos dendogramas. Por isso o intervalo a ser testado para todas as décadas será de 2 a 6 *clusters*.

Na Tabela 2.5 é trazido o valor do SI para o indicador *Crescimento do PIB* nas diferentes décadas.

Tabela 2.5 - SI para *k-means* – Indicador Crescimento do PIB

Década	**Núm. observações**	**Núm. *clusters* analisados**	**SI**	***k* recomendado**
1965-1974	124	[2 3 4 5 6]	[0,6875 0,7072 0,7138 0,7537 0,6956]	5
1975-1984	158	[2 3 4 5 6]	[0,6669 0,6678 0,7124 0,6781 0,6908]	4
1985-1994	189	[2 3 4 5 6]	[0,8781 0,6342 0,6717 0,6479 0,6484]	2
1995-2004	204	[2 3 4 5 6]	[0,7384 0,6387 0,6950 0,6982 0,7433]	6
2005-2014	204	[2 3 4 5 6]	[0,7104 0,6642 0,6281 0,6856 0,7008]	2

Ao longo do período o *k* recomendado varia entre 2 e 6, o que indica que o indicador sofre mudança significativas ao longo do tempo. . Na Tabela 2.6, onde são apresentados os dados dos *clusters* formados, posicionam-se último lugar países de instabilidade econômica ou política, e que apresentaram índices muito baixo de crescimento do PIB ou ainda negativos. Na última década houve um recuo no valor deste indicador quando comparado a década anterior, provavelmente ocasionada pela crise dos anos 2008-2009 que se iniciou nos Estados Unidos e que acabou afetando economias do mundo todo. Nas mais altas taxas de crescimento econômico se destacam países do Oriente Médio, com economia baseada no petróleo, cujo valor vem aumentando desde os anos 2000, impactando positivamente no indicador destes países.

Tabela 2.6 - Dados dos *clusters* formados com *k* recomendado – Indicador Crescimento do PIB (em %)

Ind. Crescimento do PIB	**1965-1974**	**1975-1984**	**1985-1994**	**1995-2004**	**2005-2014**
Total de países	124	158	189	204	204
Qtde países *Cluster* 1	8	56	16	31	119
Qtde países *Cluster* 2	32	65	173	57	85
Qtde países *Cluster* 3	25	16	-	13	-
Qtde países *Cluster* 4	28	21	-	95	-
Qtde países *Cluster* 5	31	-	-	1	
Qtde países *Cluster* 6	-	-	-	7	
Valor máximo	18,633	14,118	17,338	24,719	13,038
Valor mínimo	0,253	-6,699	-21,150	-1,323	-3,654
Mediana	4,999	3,427	3,175	3,908	3,774
País do vlr máx	Arábia Saudita	Omã	Eritreia	Guiné Equatorial	Catar
País do vlr mín	Níger	Moçambique	Azerbaijão	Congo	Aruba
País mediana	França	Bulgária	Jordânia	Austrália	Honduras
Exemplo país *cluster* 1'	Irã	República Dominicana	Azerbaijão	Coreia	França
Exemplo país *cluster* 2'	Equador	Mauritânia	Uganda	Venezuela	Bolívia
Exemplo país *cluster* 3'	República Centro-Africana	Gana	-	Ilhas Marshall	-
Exemplo país *cluster* 4'	Tailândia	Coreia	-	Tanzânia	-
Exemplo país *cluster* 5'	Guiana			Guiné Equatorial	
Exemplo país *cluster* 6'	-			Maldivas	

Os dendogramas gerados para o indicador *Inflação dos preços ao consumidor* são mostrados nas Figuras 2.17 a 2.21.

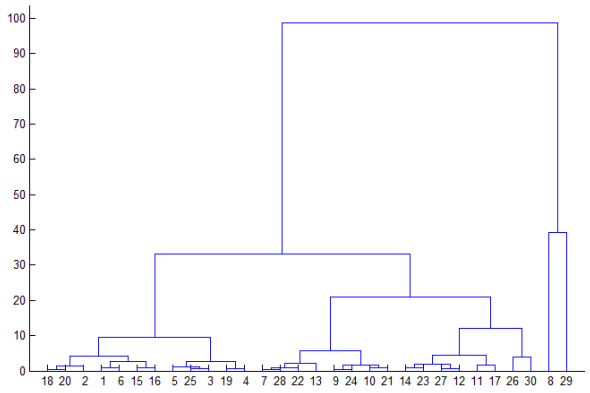

Figura 2.17: Dendograma para Inflação (1965-1974)

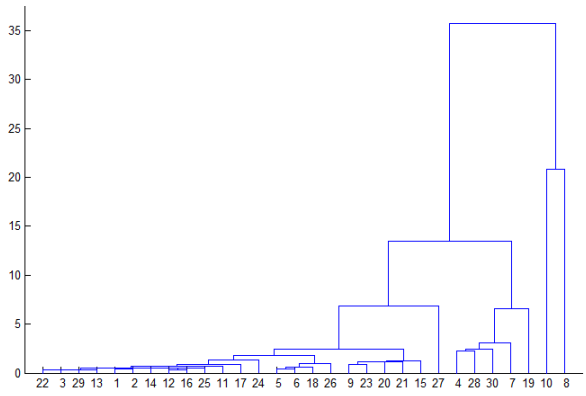

Figura 2.18: Dendograma para Inflação (1975-1984)

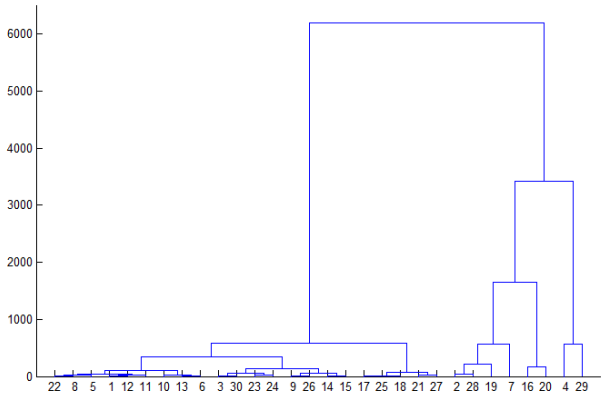

Figura 2.19: Dendograma para Inflação (1985-1994)

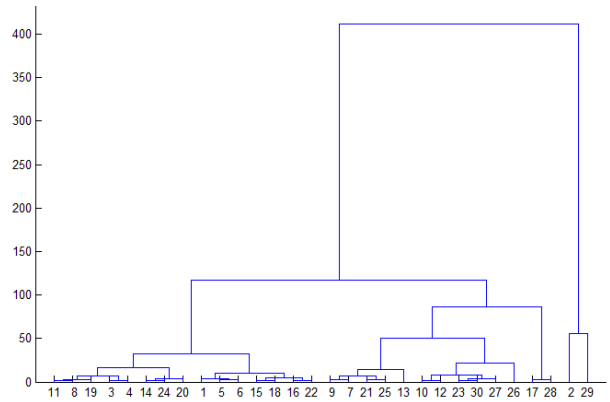

Figura 2.20: Dendograma para Inflação (1995-2004)

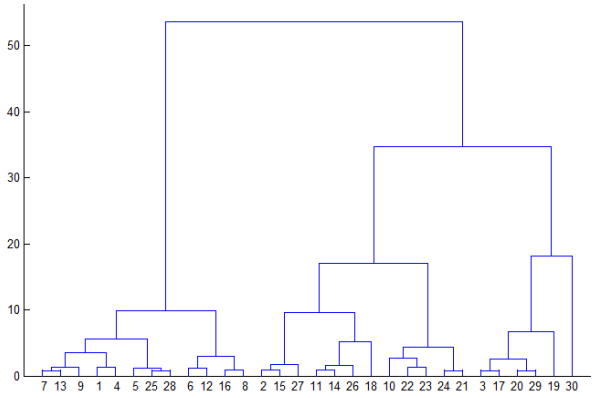

Figura 2.21: Dendograma para Inflação (2005-2014)

Nos dendogramas do indicador *Inflação* é possível observar que nesse indicador há presença de um maior número de grupos e menores distâncias entre eles, quando comparado aos indicadores anteriores, dando destaque para as décadas de 1975 a 1994. No dendograma 2.21 se identifica mais *clusters* distintos entre si. Será então testado intervalo de 2 a 6 para as três primeiras décadas, de 2 a 4 para a década de 1995 a 2004 e até 8 *clusters* para a última década.

Na Tabela 2.7 apresenta o valor do SI para o indicador de inflação nas diferentes décadas. Dos indicadores analisados, é o que possui menor variação ao longo do período, mantendo basicamente dois *clusters* para os países.

Tabela 2.7 - SI para *k-means* – Indicador Inflação

Década	**Núm. observações**	**Núm. *clusters* analisados**	**SI**	**k recomendado**
1965-1974	98	[2 3 4 5 6]	[0,9748 0,8248 0,7230 0,7607 0,6822]	2
1975-1984	118	[2 3 4 5 6]	[0,9529 0,9511 0,8700 0,6892 0,6729]	2
1985-1994	154	[2 3 4 5 6]	[0,9842 0,9757 0,9825 0,9318 0,9300]	2
1995-2004	173	[2 3 4]	[0,9949 0,9199 0,8289]	2
2005-2014	185	[2 3 4 5 6 7 8]	[0,7387 0,7824 0,7280 0,7724 0,6803 0,6486 0,6692]	3

A partir do *k* recomendado encontrado, os países foram agrupados em *clusters*, sendo a grande parcela dos países alocados a um deles, e outro *cluster* formado por uma minoria. A Tabela 2.8 resume as informações dos *clusters* formados para o indicador *Inflação*.

Tabela 2.8 - Dados dos *clusters* formados com *k* recomendado – Indicador Inflação (em %)

Indicador Inflação	**1965-1974**	**1975-1984**	**1985-1994**	**1995-2004**	**2005-2014**
Total de países	98	118	154	173	185
Qtde países *Cluster* 1	2	111	150	171	107
Qtde países *Cluster* 2	96	7	4	2	15
Qtde países *Cluster* 3	-	-	-	-	63
Valor máximo	74,973	118,376	3373,474	241,705	27,634
Valor mínimo	0,808	1,8846	-2,118	-1,069	0,0122
Mediana	4,662	10,117	8,797	4,137	3,964
País do vlr máx	Uruguai	Brasil	Armênia	Congo	Venezuela
País do vlr mín	Etiópia	Djibuti	Níger	Kosovo	Japão
País mediana	Itália	Bahrein	Ásia Meridional	Comores	Letônia
Exemplo país *cluster* 1'	Indonésia	Etiópia	Índia	Costa do Marfim	Grécia
Exemplo país *cluster* 2'	Irlanda	Israel	Armênia	Angola	Gana
Exemplo país *cluster* 3'	-	-	-	-	Trinidad e Tobago

Em todo o período é possível verificar que países com relativa instabilidade política, representados basicamente por latino-americanos, apontam em primeiro lugar considerando a maior taxa de inflação para posicioná-los.

Observando a última década, é possível identificar que os países com as menores taxas de inflação anual estão alocados no *cluster* 1, os de taxa média ao *cluster* 3 e os de maiores taxas ao *cluster* 2.

2.5 Conclusões

A clusterização se mostrou uma técnica adequada para indicadores econômicos. Apesar de o contexto econômico e social de cada país não fazer parte da base de dados, os *clusters* formados possuem distinção uns dos outros, já que foram gerados a partir do SI.

O fato de analisar cada década separadamente possibilitou verificar a evolução dos indicadores, além de identificar países que mantêm ou variam determinado desempenho ao longo do período.

Os países que formam os *clusters* com melhores índices em cada um dos indicadores devem ser analisados em detalhe para então basearem diretrizes que orientem os países de desempenho relativamente pior.

Como sugestão para futuros estudos, recomenda-se substituir os dados de mediana das décadas pelos valores reais (em escala anual) no procedimento de clusterização. podendo também ser aplicada a clusterização usando vetores temporais. Ou ainda, para uma análise mais generalista, poderia ser testada a inclusão de dados de diferentes indicadores numa mesma base e verificar a clusterização resultante. Os agrupamentos podem ser utilizados para apontar alternativas de ações que promovam incrementos no desempenho dos países nos indicadores observados.

2.6 Referências

ALMEIDA, P. R. *A economia internacional no século XX: um ensaio de síntese.* Revista Brasileira de Política Internacional, vol.44, no.1, Jan./Jun, 2001.

ANZANELLO, M. J.; FOGLIATTO, F. S. *Selecting the best clustering variables for grouping mass-customized products involving workers' learning*. International Journal of Production Economics, v. 130, n. 2, p. 268-276, 2011.

APERGIS, N.; CHRISTOU, C.; MILLER, S. *Convergence patterns in financial development: evidence from club convergence*. Empirical Economics, v. 43, n. 3, p. 1011-1040, 2012.

AUSLOOS, M; LAMBIOTTE, R. *Clusters or networks of economies? A macroeconomy study through Gross Domestic Product*. Physica A: Statistical Mechanics and its applications, v. 382, n. 1, p. 16-21, 2007.

BATTISTI, M.; PARMETER, C. F. *Clustering and polarization in the distribution of output: A multivariate perspective*. Journal of Macroeconomics, v. 35, p. 144-162, 2013.

CAMACHO, M.; PEREZ-QUIROS, G.; SAIZ, L. *Do European business cycles look like one?*. Journal of Economic Dynamics and Control, v. 32, n. 7, p. 2165-2190, 2008.

CRUZ, M. D.; OCHI, L. S. *Um Algoritmo Evolutivo com Memória Adaptativa para o Problema de Clusterização Automática*. Learning and Nonlinear Models, v. 8, n. 4, p. 227-239, 2011.

DE ALENCAR, B. J.; BARROSO, L. C.; DE ABREU, J. F. *Análise Multivariada de Dados no Tratamento da Informação Espacial: uma abordagem com análise de Agrupamentos*. Revista Iberoamericana de Sistemas, Cibernética e Informática, v. 10, n. 2, p. 6-12, 2013.

DE LA PAZ-MARÍN, M.; GUTIÉRREZ, P. A.; HERVÁS-MARTÍNEZ, C. *Classification of countries' progress toward a Knowledge Economy based on machine learning classification techniques*. Expert Systems with Applications, v. 42, n. 1, p. 562-572, 2015.

FIORUCI, J. A.; TOLEDO, F. M.B.; NASCIMENTO, M. C. V. *Heuristics for minimizing the maximum within-clusters distance*. Pesqui. Oper., Rio de Janeiro , v. 32, n. 3, p. 497-522, nov. 2012.

GIL, A. C. *Como classificar as pesquisas*. _____. Como elaborar projetos de pesquisa, v. 4, p. 41-56, 2002.

HARTIGAN, J. A. *Statistical theory in clustering*. Journal of classification, v. 2, n. 1, p. 63-76, 1985.

HENNIG, C. *Clusters, outliers, and regression: fixed point clusters*. Journal of Multivariate Analysis, v. 86, n. 1, p. 183-212, 2003.

HENNIG, C.; LIAO, T. F. *How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification.* Journal of the Royal Statistical Society: Series C (Applied Statistics), v. 62, n. 3, p. 309-369, 2013.

JOHNSON, R.A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 3 ed. Prentice-Hall International (1992).

KUBRUSLY, L. S. *Um procedimento para calcular índices a partir de uma base de dados multivariados*. Pesqui. Oper., Rio de Janeiro , v. 21, n. 1, p. 107-117, jun. 2001 .

LEE, P. M. *Use Of Data Mining In Business Analytics To Support Business Competitiveness*. Review of Business Information Systems (RBIS), v. 17, n. 2, p. 53-58, 2013.

LENGYEL, I.; RECHNITZER, J. *The competitiveness of regions in the Central European transition countries*. Macrotheme Rev, v. 2, p. 106-121, 2013.

LÓPEZ-VILLAVICENCIO, A; MIGNON, V. *On the impact of inflation on output growth: Does the level of inflation matter?*. Journal of Macroeconomics, v. 33, n. 3, p. 455-464, 2011.

LIU, Y. et al. *Understanding of internal clustering validation measures*. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010. p. 911-916.

MINGOTI, S. A.; LIMA, J. O. *Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms*.European Journal of Operational Research, v. 174, n. 3, p. 1742-1759, 2006.

MONTANARI, M. G.; MIRANDA, R. A. M.; GIRALDI, J. M. E. *Uma comparação quantitativa entre dois índices de competitividade*. Revista de Globalización, Competitividad y Gobernabilidad, v. 8, n. 2, p. 121, 2014.

OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. *Clusterização em Mineração de Dados*. Instituto de Computação-Universidade Federal Fluminense-Niterói, 2004.

PHILLIPS, P. C. B; SUL, D. *Transition modeling and econometric convergence tests*. Econometrica, v. 75, n. 6, p. 1771-1855, 2007.

POWELL, W.; SNELLMAN, K. *The knowledge economy*. Annual review of sociology, p. 199-220, 2004.

RENCHER, A. *Methods of Multivariate Analysis*, Wiley-Interscience (2002).

RESENDE, M. F. C.; VIEIRA, F. A. C. *Taxa de câmbio real e déficits gêmeos na economia monetária: relações de causalidade*. ECONOMIA & TECNOLOGIA, 2011.

ROUSEEUW, P. J. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics. Vol. 20, No. 1, 1987, pp. 53–65.

Sítio do Banco Mundial (http://www.worldbank.org/). Acessado em: outubro de 2015.

SILVA, J. G. S. et al. *Uma Experiência Didática com Base no Emprego do Matlab nos Cursos de Graduação da Faculdade de Engenharia da UERJ*. In: World Congress on Engineering and Technology Education-WCETE. 2004. p. 1-5.

SNIEŠKA, V. *Research into international competitiveness in 2000–2008*. Engineering Economics, v. 59, n. 4, 2015.

SPERO, J. E., HART, J. *The Politics of International Economic Relations*, Wadsworth (2010).

STEVANS, L. K. et al. *The Economic Competitiveness of Countries: A Principal Factors Approach*. International Journal of Economics and Finance, v. 4, n. 12, p. p76, 2012.

VYAS, S; KUMARANAYAKE, L. *Constructing socio-economic status indices: how to use principal components analysis*. Health policy and planning, v. 21, n. 6, p. 459-468, 2006.

3 Segundo Artigo: Sistemática para seleção de indicadores econômicos com vistas ao agrupamento de países

Paula Lunardi de Mello

Michel José Anzanello

Resumo

Indicadores que descrevem as principais dimensões econômicas de um país podem ser utilizados como norteadores na elaboração e monitoramento de políticas de desenvolvimento e crescimento desses países. Neste sentido, a aplicação de técnicas de análise multivariada de dados pode auxiliar no desafio de se tratar esse banco de dados, gerando medidas e ações apropriadas. Este artigo apoia-se em dados do Banco Mundial para identificar os indicadores mais relevantes na formação de agrupamentos de países com perfis similares de desenvolvimento econômico. Para tanto, integra a técnica de clusterização do tipo *k-means* ao procedimento "*leave one variable out at a time*" de seleção de variáveis (indicadores), avaliando a qualidade dos *clusters* gerados através do *Silhouette Index* (SI). Partindo da análise de 10 indicadores de cunho econômico que descrevem 84 países, apenas três ficam retidos ao final da aplicação da sistemática proposta, gerando resultado satisfatório quanto à representatividade e qualidade da clusterização gerada.

Palavras-chave: indicadores econômicos, clusterização, seleção de variáveis, *silhouette index*.

3.1 Introdução

O volume de dados coletados nas mais diversas frentes e aplicações tem crescido sistematicamente com a evolução dos recursos computacionais. A *Business Software Alliance* (BSA), que representa algumas das maiores empresas de software da atualidade (Apple, Microsoft, IBM e Intel, dentre outras), realizou estudo em 2015 apontando que cerca de 2,5 quintilhões de bytes são criados diariamente. Esse volume de dados se refere aos mais diferentes fenômenos e processos, sendo uma boa fatia gerada e armazenada no ambiente das organizações empresariais e governamentais. Para a BSA, os dados precisam ser coletados, armazenados, analisados e convertidos para que se chegue às informações apoiadoras da

tomada de decisão. Assim é possível relacionar essa necessidade a alguns conceitos atuais para o gerenciamento de dados, como *Big Data*, por exemplo.

Breternitz & Silva (2013) definem *Big Data* como um conjunto de tendências tecnológicas que, através de uma nova abordagem, permite tratar e entender essas grandes bases de dados para dar apoio à tomada de decisão. Einav & Levin (2014) afirmam que, ao longo das próximas décadas, o *Big Data* mudará o panorama da política econômica e da pesquisa em economia, não substituindo, no entanto, o senso comum, a teoria econômica ou a necessidade de projetos de pesquisa, mas os complementando. Ainda segundo o estudo da BSA, a economia movida a dados pode causar impactos em diversos setores, como manufatura, transporte, energia, agricultura, educação e saúde. Neste estudo, os economistas estimam que ganhos de eficiência gerados por meio do adequado processamento dessas informações podem agregar US$ 15 trilhões ao PIB global até 2030.

Indicadores econômicos de desempenho de países oferecem um vasto histórico do perfil financeiro daqueles países, podendo ser utilizados como norteadores para que políticas de desenvolvimento e crescimento sejam elaboradas. Neste sentido, a aplicação de técnicas de análise multivariada de dados pode auxiliar no desafio de se tratar esse banco de dados, gerando medidas e ações apropriadas. No entanto, o custo do gerenciamento de um volumoso banco pode inviabilizar o tratamento de dados, justificando a busca por modelos compostos por um número reduzido de variáveis, os quais mostram uma alternativa eficiente nessa minimização de custos e esforço computacional (ZIMMER & ANZANELLO, 2014).

Este artigo utiliza dados do Banco Mundial para identificar os indicadores mais relevantes na formação de agrupamentos de países com perfis similares de desenvolvimento econômico. Para tanto, integra a técnica de clusterização do tipo *k-means* ao procedimento "*leave one variable out at a time*" de seleção de variáveis (indicadores). Neste procedimento iterativo de seleção de variáveis, cada indicador é omitido momentaneamente do banco de dados e uma clusterização é rodada; a qualidade do *cluster* de gerado é avaliada pelo *Silhouette Index* (SI). Na sequência, tal indicador é devolvido ao banco e o indicador subsequente é omitido temporariamente. Depois que todos os indicadores são omitidos, o indicador faltante que conduziu ao máximo SI é removido em definitivo, partindo-se do pressuposto que tal indicador é o menos relevante para a formação dos agrupamentos. O procedimento é repetido para as variáveis remanescentes, até que sobre apenas uma variável.

O subconjunto de variáveis conduzindo ao máximo SI global é então recomendado para clusterizações futuras.

Em termos práticos, a formação de agrupamentos de países com perfis similares de indicadores permite aprofundar o entendimento de fatores implícitos que justificam comportamentos semelhantes no âmbito econômico, permitindo ainda que medidas bem sucedidas em alguns países sejam estendidas a outros países pertencentes àquele *cluster* (devido às características similares).

O trabalho está dividido em cinco seções. A seção 3.2 apresenta a revisão da literatura no que diz respeito a ferramentas de clusterização de dados, aprofundando a pesquisa em métodos para a seleção de variáveis. A seção 3.3 descreve o método aplicado para geração dos *clusters* e seleção das variáveis mais relevantes através da técnica "*leave one variable out at a time*". Já na seção 3.4 são mostrados os resultados obtidos pela aplicação do método aos dados. Por fim, na seção 3.5 são expostas as conclusões geradas pelo estudo.

3.2 Revisão da literatura

3.2.1 Ferramentas de clusterização de dados

A clusterização é uma importante técnica usada para dividir elementos de dados em subconjuntos homogêneos (chamados *clusters*), dentro dos quais os elementos são mais semelhantes uns aos outros, enquanto são mais diferentes em relação aos elementos de outros grupos (STARCZEWSKI & KRZYŻAK, 2015). Johnson & Wichern (1992) afirmam que a clusterização realiza o agrupamento de itens em função das similaridades ou distâncias (dissimilaridades) entre observações.

Como citado por Wolfson et al. (2004), para realizar uma clusterização, decisões básicas devem ser tomadas quanto ao procedimento e quanto ao método (aglomerativo ou divisivo) a ser utilizado. Os procedimentos se dividem hierárquicos e não-hierárquicos (HAIR et al., 2005). O hierárquico parte da construção de estrutura do tipo árvore, que por sua vez tem também uma distinção entre métodos aglomerativos e divisivos (HAIR et al., 2005). Nos métodos aglomerativos, começa-se com cada observação sendo um único *cluster*; nos passos seguintes, os dois grupos mais próximos são combinados em um novo, reduzindo o número de agrupamentos em uma unidade a cada passo (HAIR et al., 2005).

De forma contrária, o método divisivo começa com um único *cluster*; nos passos sucessivos, as observações mais diferentes entre si se separam e são transformadas em outros grupos menores (WOLFSON et al., 2004; HAIR et al., 2005). Johnson e Wichern (1992) esclarecem que o resultado de ambos os métodos, aglomerativo e divisivo, pode ser representado em um gráfico de duas dimensões, chamado de dendograma, sendo possível identificar as junções e divisões que foram realizadas nos diferentes níveis, como ilustra a Figura 3.1. Segundo Wolfson et al. (2004), o dendograma pode ser definido como um diagrama de árvore que mostra a clusterização aglomerativa, onde a base da árvore tem todas as observações e, à medida que se move para direita da árvore, estes casos individuais se combinam em grupos com base no número de etapas necessárias para formá-los usando o método de aglomeração selecionado.

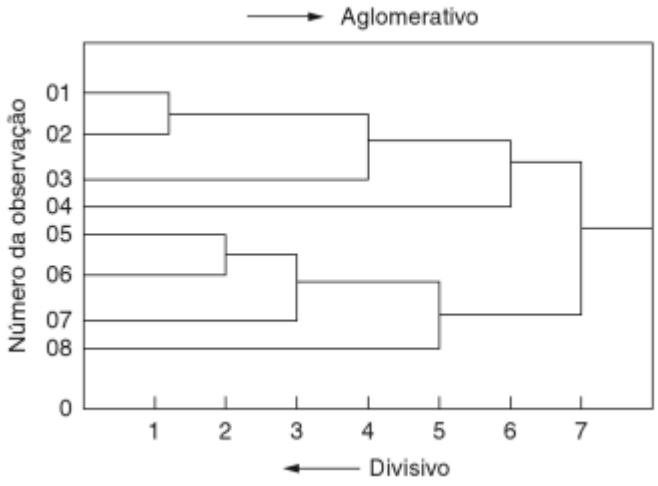

Figura 3.1: Dendograma ilustrando agrupamento hierárquico (HAIR et al., 2005)

Em Wolfson et al. (2004), a lógica de cada um dos métodos aglomerativos é resumida em: a) ligação única: mede a distância entre os *clusters* através da distância entre os dois pontos nos *clusters* mais próximos uns dos outros; b) ligação completa: mede a distância entre *clusters* através da distância entre os dois pontos mais afastados entre eles; c) ligação centroide: tenta determinar o "centro" do agrupamento; d) ligação média: utiliza a distância média entre todos os pares possíveis de pontos; e) método de Ward: baseado em uma aproximação de soma de quadrados, o qual tende a criar *clusters* de tamanho similar.

Os procedimentos não-hierárquicos, por sua vez, não envolvem o processo de construção em árvore, pois partem de uma semente de agrupamento como centro inicial de um agregado, e todas as observações dentro de uma distância de referência pré-especificada são incluídas em um mesmo agrupamento (HAIR et al., 2005). A crescente utilização desses procedimentos se justifica por serem menos suscetíveis a observações atípicas presentes nos dados.

O método *k-means* é o mais comum deles, pela forma eficaz e rápida na sua aplicação (YEDLA et al., 2010). O *k-means* parte de um número inicial de *k* grupos e agrupa cada observação ao centroide (valor médio das observações contidas no agrupamento) mais próximo (JOHNSON & WICHERN, 1992). Este método tipicamente utiliza a distância Euclidiana para identificar diferentes *clusters* nos dados (SUN et al., 2012; LEE, 2013), embora outras possam ser utilizadas.

A distância euclidiana é utilizada para avaliar a similaridade entre observações (RENCHER, 2002). Considere duas observações (objetos a serem clusterizados) descritas pelos vetores $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_p)$, ambos consistindo de *p* variáveis. A distância Euclidiana entre eles é apresentada na Eq. (3.1), sendo que a soma das distâncias euclidianas entre elementos classificados no mesmo agrupamento deve ser minimizada.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} = \sqrt{\sum_{j=1}^{p} (x_j - y_i)^2} \tag{3.1}$$

Para validação da clusterização encontrada, comumente são usados índices que medem a qualidade da solução resultante. Wang et al. (2009) dividem os tipos de índices entre internos e externos. Os índices internos compararam a estrutura de grupos descoberta com uma estrutura de grupos previamente conhecida, e nos internos a solução é avaliada usando parâmetros inerentes ao conjunto de dados, como compacidade e/ou separabilidade (DUDOIT et al., 2002). A compacidade mede, com base na variância ou distância, o quão proximamente estão relacionadas às observações de um *cluster*, enquanto a separabilidade mede o quão distintos ou bem separados um *cluster* é em relação a outros *clusters* (LIU et al., 2010).

Um índice amplamente aplicado é o S*ilhouette Index* (SI), que se baseia nas diferenças entre as distâncias das observações, analisadas aos pares dentro do *cluster* (LIU et al., 2010). Esse índice varia de -1 a +1 e o número ótimo de *clusters* é determinado pela maximização

desse índice, que mede o quão distintos estão uns dos outros os *clusters* formados. Conforme Anzanello e Fogliatto (2011), o SI_n é calculado de acordo com a Eq. (3.2):

$$SI_n = \frac{b(n) - a(n)}{\max\{b(n), a(n)\}} \tag{3.2}$$

sendo *a(n)* a média das distâncias da *n*-ésima observação dentro do mesmo *cluster*, e *b(n)* a média das distâncias da mesma observação em relação as outras alocadas no *cluster* mais próximo.

O SI não depende da técnica de clusterização utilizada na análise, uma vez que é calculado com base em distâncias. Normalmente é usado para medir a qualidade global da clusterização através da média de SI_n, conforme Eq. (3.3):

$$\overline{SI} = \frac{\sum_{n=1}^{N} SI_n}{N} \tag{3.3}$$

onde *n* o índice da observação e *N* o total de observações.

Sun et al. (2012) colocam que, enquanto o método *k-means* é simples e computacionalmente eficiente, seu desempenho pode ser deteriorado significativamente quando o *cluster* possui grande número de variáveis, que muitas vezes podem não conter informações sobre a estrutura de agrupamento e até tornar a análise muito complexa. Para superar estas dificuldades em agrupamento de dados, faz-se fundamental identificar as variáveis com maior poder de separação dos dados (JAIN, 2010; WITTEN & TIBSHIRANI, 2010; SUN et al., 2012).

Huang et al. (2005) ainda ponderam que dezenas ou centenas de variáveis são normalmente derivadas do banco de dados na seleção inicial, formando um grande espaço dimensional. É conhecido que uma estrutura de agrupamento de interesse geralmente ocorre em um subespaço definido por um subconjunto das variáveis inicialmente selecionadas, sendo importante identificar o subconjunto de variáveis relevantes a fim de encontrar a estrutura de agrupamento (HUANG et al., 2005; RAFTERY & DEAN, 2006).

3.2.2 Métodos para seleção de variáveis

A seleção de variáveis se mostra uma etapa importante na análise de agrupamento de dados em diversos segmentos e aplicações, devido à necessidade de identificar as variáveis mais relevantes para uma determinada classificação (SOTOCA & PLA, 2010). Os autores definem como relevantes as variáveis que influenciam ou determinam um conjunto de dados, de forma a minimizar o erro na classificação desse agrupamento.

Diversos benefícios emergem da seleção de variáveis em contextos de clusterização. Para Guyon & Elisseeff (2003), os principais ganhos incluem facilidade na visualização das estruturas dos dados, compreensão dos agrupamentos formados e redução dos requisitos de medição e armazenamento. Anzanello & Fogliatto (2011) corroboram a ideia, apontando que a qualidade do agrupamento depende de um conjunto limitado de variáveis relevantes.

O desempenho de agrupamento pode ser avaliado através do SI, que contribui na indicação da variável que potencialmente pode ser desconsiderada na análise (ANZANELLO & FOGLIATTO, 2011). Starczewski & Krzyżak (2015) avaliaram o desempenho do SI aplicando diferentes métodos hierárquicos para o agrupamento de dados, sendo que os resultados indicaram bons resultados do índice e permitiram identificar as variáveis mais relevantes na constituição de bons agrupamentos.

A classificação de países se mostra importante no amparo de decisões político-econômicas. Nesse sentido, Abizadeh e Basilevsky (1986) desenvolveram uma técnica para classificação de países através variáveis socioeconômicas pré-selecionadas usando método da máxima verossimilhança. Tais autores propuseram uma técnica isenta de algumas deficiências estatísticas atribuídas aos métodos mais antigos.

Berlage e Terweduwe (1988) compararam as classificações de países realizadas por agências internacionais com a classificação realizada através de clusterização e análise fatorial. Foi possível verificar que o agrupamento usado pelas agências possui viés arbitrário e pragmático, não seguindo uma metodologia técnica e robusta, o que pode minar os resultados de classificação e sua interpretação (BERLAGE & TERWEDUWE, 1988).

Por sua vez, Lletti et al. (2004) levantam o problema derivado da análise de variáveis

não-informativas, as quais tipicamente estão não diretamente relacionadas com a estrutura interna dos dados ou trazem elevados níveis de ruído. No estudo realizado, foram conduzidas simulações acerca do SI, observando-se que, à medida que variáveis não relevantes (ou não-informativas) eram incluídas na base de dados, ocorria uma degradação do SI (além de tornar mais difícil para o algoritmo *k-means* encontrar a estrutura de *cluster* adequada aos dados). Como regra geral, os autores apontam que SI's abaixo de 0,25 indicam que não há estrutura nos *clusters* encontrados. Neste sentido, os experimentos de Lletti et al. (2004) sugerem que quando mais de cinco variáveis não informativas são adicionadas à base de dados, verifica-se perda total da estrutura de agrupamento dos dados.

Raftery e Dean (2006) propuseram um método para seleção de fatores baseado em modelo, onde todo o conjunto de dados, reais e simulados, são clusterizados usando fatores aproximados de Bayes e um algoritmo guloso (espécie de algoritmo de otimização). Nos exemplos, o método forneceu melhores estimativas do número de *clusters* e taxas de erro de classificação mais baixas. Com propósitos semelhantes, Steinley e Brusco (2007) sugeriram um procedimento de ponderação de variável operando conjuntamente com a seleção de variáveis; tal sistemática de ponderação é fundamentada na variabilidade inerente aos dados que exibem estrutura de *cluster*. O método foi aplicado na base de dados multivariados de Fisher e vários conjuntos de dados sintéticos, e Steinley e Brusco (2007) chegaram a resultados favoráveis quando comparados com os métodos de padronização existentes.

Steinley e Brusco (2008) avaliaram ainda oito diferentes técnicas de seleção oriundas de diferentes áreas, como estatística, aprendizagem de máquina e psicologia. O método de melhor desempenho foi o proposto em Steinley e Brusco (2007), enquanto que os de pior foram os baseados em modelos de mistura finita[1].

Em estudo recente, Cervo & Anzanello (2015) propuseram uma sistemática para seleção das variáveis de clusterização mais relevantes com base em índices de importância de variáveis, avaliando variações no SI médio de acordo com a porcentagem de variáveis retidas. O índice de importância de variáveis apoiava-se em parâmetros oriundos da Análise de

[1] A técnica de mistura finita modela os dados como uma soma ponderada de g densidades de probabilidade, cada qual representando uma subpopulação, cujos pesos representam a força daquele grupo (PINTO, 2006).

Componentes Principais.

3.3 Método

Esta pesquisa pode ser classificada como de natureza aplicada, utilizando abordagem quantitativa, pois analisa dados reais da economia dos países em questão, focada em dados temporais. Seu objetivo é exploratório, por buscar proporcionar maior familiaridade com o tema (GIL, 2002).

O estudo foi realizado utilizando dados do Banco Mundial reunidos pelo *Global Federation of Competitiveness Councils* (GFCC). Essa federação é uma rede de líderes de conselhos de competitividade em nível mundial e define direcionadores (*drivers*) fundamentais da competitividade nacional que sustentam o crescimento econômico global. Esses *drivers* são: desempenho geral, complexidade econômica, infraestrutura, talento, capital, inovação, qualidade de vida e crescimento futuro. Para a análise deste artigo, foram escolhidos dez indicadores econômicos agrupados pertencentes ao *driver* desempenho geral, apresentados no Anexo A.

Os indicadores selecionados para este trabalho foram todos aqueles que na base de dados possuíam levantamento de 1995 a 2014, contemplando todos os países com informações no intervalo.

A sistemática de seleção de indicadores para aprimoramento da clusterização foi aplicada a um banco de dados composto por 84 países e dez indicadores. Foi calculada a mediana dos países para cada indicador, por ser uma estatística menos suscetível a valores discrepantes. Para a construção dos dendogramas, foi utilizada a distância euclidiana e o método de ligação de Ward, já aplicados em pesquisas anteriores relacionadas a indicadores econômicos (KUBRUSLY, 2001; MONTANARI et al., 2014). Os procedimentos computacionais foram realizados no software Matlab; como citado por Silva (2004), tal aplicativo é amplamente utilizado para análise de dados quantitativos. A função utilizada foi a *evalclusters*, método *k-means*.

Os dez indicadores foram colocados em uma mesma base de dados e as medianas do intervalo de 1995 a 2014 foi calculada para cada um dos países, conforme mostrado na tabela do Anexo B.

Usando o método de ligação de Ward, foi gerado o dendograma com os dados de todos os indicadores. Conforme Wolfson et al. (2004), uma linha de corte horizontal deve ser desenhada para definir o número de grupos de uma maneira visual.

A clusterização dos dados foi realizada pelo método *k-means* e calculados os valores de SI médio para um intervalo de *k* a partir da informação número de *clusters* do passo anterior. Na sequência, cada indicador é omitido momentaneamente do banco de dados e uma clusterização *k-means* é rodada usando distância Euclidiana. A qualidade do *cluster* formado é avaliada pelo SI, e então tal indicador é devolvido ao banco; o procedimento é repetido para o indicador subsequente até que todos os indicadores tenham sido omitidos uma vez do banco.

Depois que todos os indicadores são omitidos, o indicador que conduziu ao máximo SI é removido em definitivo partindo-se do pressuposto que tal indicador é o menos relevante para a formação dos agrupamentos, de acordo com o procedimento "*leave one variable out at a time*". Volta-se ao passo 3 até que reste um único indicador na base de dados.

3.4 Resultados e Discussões

Realizado o cálculo da mediana dos dez indicadores, foi gerado o dendograma da Figura 3.2. A maior distância - representada no eixo vertical - entre os agrupamentos se dá entre 2 e 4 *clusters*. Por esse motivo, foi esse o intervalo usado para cálculo do SI (de k=2 a k=4). Embora a legenda indicando os países não se apresente de forma visualizável na figura (bem como os agrupamentos formados nos primeiros estágios de clusterização), apresenta-se a mesma com o propósito restrito de identificar um intervalo apropriado de *clusters* a serem formados (identificável na parte superior do dendograma).

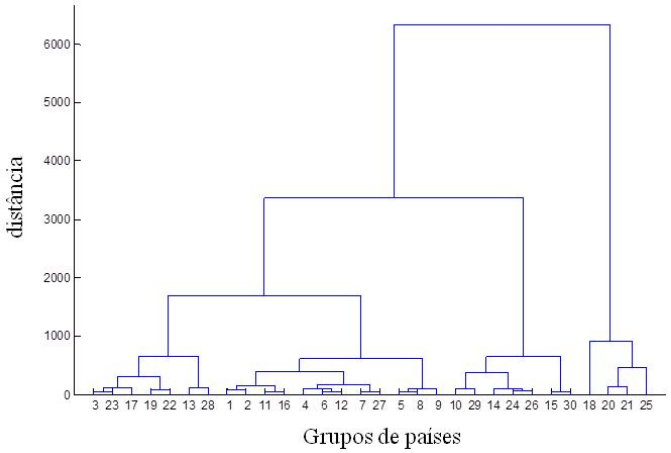

Figura 3.2: Dendograma dos dados dos 10 indicadores x 84 países

Realizando o agrupamento dos dados, foram calculados os valores de SI médio para um intervalo de dois a quatro *clusters*, com resultados representados na Tabela 3.1. A mesma tabela sugere que o número recomendado de *clusters* a ser formado é dois (SI máximo).

Tabela 3.1 - SI médio para o intervalo de *clusters*

Nº de *clusters*	**SI médio**
2	0,9871
3	0,9825
4	0,9613

Definido o nº recomendado de *clusters* como dois, foram calculados os SI's médios para a base de dados. Os resultados são sintetizados na Tabela 3.2. O resultado do SI médio quando todos os indicadores são agrupados em dois *clusters* é igual a 0,9048, que denota boa qualidade da clusterização.

Tabela 3.2 - SI médios para a clusterização dos indicadores

Iteração	Indicadores	SI Médio
1ª	I1 a I10	0,9048
2ª	- I1	0,9171
	- I2	0,905
	- I3	0,8682
	- I4	0,8973
	- I5	0,9039
	- I6	**0,9209**
	- I7	0,9043
	- I8	0,8999
	- I9	0,8979
	- I10	0,8994
3ª	**- I6 - I1**	**0,9274**
	- I6 - I2	0,921
	- I6 - I3	0,8771
	- I6 - I4	0,913
	- I6 - I5	0,9156
	- I6 – I7	0,9201
	- I6 – I8	0,9175
	- I6 – I9	0,9135
	- I6 – I10	0,9071
4ª	-I6 - I1 - I2	0,9211
	-I6 - I1 - I3	0,9134
	-I6 - I1 - I4	0,9179
	-I6 - I1 - I5	**0,941**
	-I6 - I1 - I7	0,919
	-I6 - I1 - I8	0,9314
	-I6 - I1 - I9	0,9186
	-I6 - I1 - I10	0,9381
5ª	-I6 - I1 - I5 - I2	0,9471
	-I6 - I1 - I5 - I3	0,9452
	-I6 - I1 - I5 - I4	0,932
	-I6 - I1 - I5 - I7	0,9406
	-I6 - I1 - I5 - I8	0,9438
	-I6 - I1 - I5 - I9	0,9326
	-I6 - I1 - I5 - I10	**0,9504**

Iteração	Indicadores	SI Médio
6ª	-I6 - I1 - I5 - I10 - I2	0,9585
	-I6 - I1 - I5 - I10 - I3	0,9085
	-I6 - I1 - I5 - I10 - I4	0,9407
	-I6 - I1 - I5 - I10 - I7	0,958
	-I6 - I1 - I5 - I10 - I8	**0,9594**
	-I6 - I1 - I5 - I10 - I9	0,9414
7ª	**-I6 - I1 - I5 - I10 - I8 - I2**	**0,9564**
	-I6 - I1 - I5 - I10 - I8 - I3	0,778
	-I6 - I1 - I5 - I10 - I8 - I4	0,9497
	-I6 - I1 - I5 - I10 - I8 - I7	0,9563
	-I6 - I1 - I5 - I10 - I8 - I9	0,9508
8ª	-I6 - I1 - I5 - I10 - I8 - I2 - I3	0,779
	-I6 - I1 - I5 - I10 - I8 - I2 - I4	0,9442
	-I6 - I1 - I5 - I10 - I8 - I2 – I7	**0,9628**
	-I6 - I1 - I5 - I10 - I8 - I2 - I9	0,9436
9ª	-I6 - I1 - I5 - I10 - I8 - I2 – I7 - I3	0,7323
	-I6 - I1 - I5 - I10 - I8 - I2 – I7 - I4	**0,9472**
	-I6 - I1 - I5 - I10 - I8 - I2 – I7 - I9	0,9453
10ª	-I6 - I1 - I5 - I10 - I8 - I2 – I7 - I4 - I3	0,7463
	-I6 - I1 - I5 - I10 - I8 - I2 – I7 - I4 - I9	**0,901**

Visto isso, é possível verificar que na segunda iteração do método, quando omitido o indicador *Importação de bens e serviços* (I6), o SI aumenta aproximadamente 1,6% quando comparado ao SI inicial. Em diversos estudos de indicadores econômicos, a *Importação de bens e serviços* não recebeu destaque em relação ao seu impacto na avaliação econômica de países (BARRO, 1991; ANDERTON & TEWOLDE, 2011; KHRAMOV & LEE, 2013). De tal forma, entende-se como coerente a remoção de tal indicador. Na terceira iteração, o SI foi maximizado ao ser retirado o indicador *Exportação de bens e serviços* (I1), mas o incremento no índice foi bem menor.

Houve aumento do SI na quarta iteração, e o indicador desconsiderado foi o *Formação Bruta de Capital Fixo* (I5), aumentando o SI em pouco mais de 1,3 ponto percentual. Esse indicador por definição está intimamente ligado ao somatório de riquezas disponíveis para

investimento no país, como detalhado no estudo de Kilsztajn (1998), em que o índice é calculado em função do PIB. Por ser uma variável dependente, ela pode ser desconsiderada para a iteração seguinte.

Com a retirada do indicador *População total* (I10) na quinta iteração o SI atinge o valor de 0,9504. Surpreende esse resultado por ser esta uma variável que impacta indiretamente em diversos indicadores, principalmente aqueles que medem o nível de produção e renda de países (CONSIGLIERE & CARNEIRO, 1970). Mas, seguindo a metodologia, o I10 será desconsiderado.

Na sexta o máximo SI médio resultou em 0,9594, quando omitido o indicador *Formação Bruta de Capital Fixo Nominal* (I8), ligeiramente mais alto que a anterior. Analisando sob a mesma ótica da retirada do indicador *Formação Bruta de Capital Fixo* (I5), por se referirem ao mesmo indicador expressado em diferentes unidades, o I8 é retirado.

Uma leve queda no SI pode ser observada na sétima iteração, que obteve máximo SI médio no descarte do indicador *PIB per capita* (I2). Pelo cálculo do I2 ser o resultado do quociente entre o I1 e o I10, e este último já ter sido desconsiderado nas etapas anteriores, o impacto de I1 para a clusterização parece maior que I2 e I10. Desta forma, a sistemática segue descartando I2.

O SI cresce novamente com a retirada do *Indicador Rendimento Nacional Bruto per capita com base na paridade do poder de compra* (I7) na oitava iteração. Consigliere e Carneiro (1970) mostram em seu estudo que o RNB é influenciado pelo tamanho da população, formação bruta de capital fixo, exportações e importações. Como esses indicadores já mostraram pouca influência no agrupamento nos passos anteriores, é esperado que o RNB também seja descartado. Nas iterações subsequentes, o SI máximo volta a cair quando comparado às iterações preliminares, chegando a um valor próximo ao inicial.

De tal forma, percebe-se que os indicadores retidos que maximizaram o SI foram *PIB* (I3), *crescimento PIB* (I4) e *Inflação* (I9), reafirmando sua representatividade na caracterização de agrupamento de indicadores econômicos de países (AUSLOOS & LAMBIOTTE, 2007; LÓPEZ-VILLAVICENCIO & MIGNON, 2011), figurando também como importantes indicadores de desempenho da economia do país (KHRAMOV & LEE,

2013). A evolução do SI pode ser observada na Figura 3.3, mostrando que seu valor máximo é atingido quando apenas três indicadores ficam retidos.

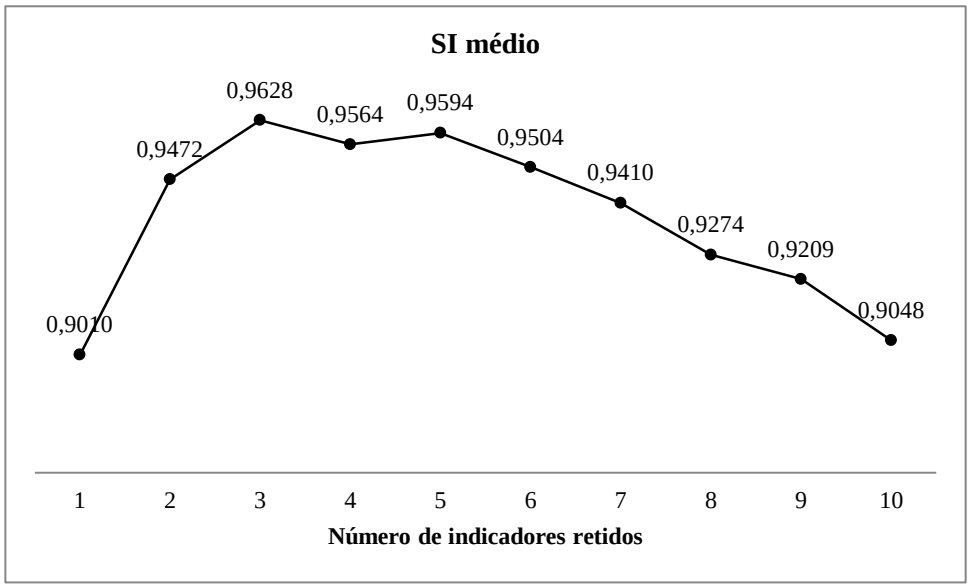

Figura 3.3: SI médio de acordo com número de indicadores retidos

3.5 Conclusões

A análise de desempenho de países pode ser auxiliada através da clusterização de seus indicadores econômicos. Nas proposições deste artigo, técnicas de agrupamento de países e de seleção de variáveis se mostraram adequadas e trouxeram informações relevantes a respeito dos dados. Percebe-se consistência na clusterização realizada, também mostrada na análise das distâncias entre os *clusters*, conforme dendograma da Figura 3.2, onde a distância entre os agrupamentos fica em destaque.

Em relação ao desempenho de diferentes países, um índice que pode ser aplicado é o Índice de Desempenho Econômico (IDE). Ele foi gerado pelo Fórum Monetário Internacional e utiliza variáveis reduzidas e estrutura simples (KHRAMOV & LEE, 2013). É calculado com base na taxa de inflação, na taxa de desemprego, taxa de déficit público em relação ao PIB e taxa de crescimento do PIB. Khramov e Lee (2013) definem quais seriam as taxas recomentadas para cada um dos índices, através do estudo de teorias econômicas relatadas em seu relatório. Dessa maneira, o resultado encontrado (variáveis retidas foram *PIB* (I3), *crescimento PIB* (I4) e *Inflação* (I9)), reforçam-se as análises realizadas por Khramov e Lee

(2013), que definem esses três indicadores como os que melhor representam o desempenho econômico de países, juntamente com a taxa de desemprego e a taxa de déficit publico, pois todos compõem a equação do índice de desempenho econômico. O Produto Interno Bruto (PIB) indica o nível de riqueza do país (AUSLOOS & LAMBIOTTE, 2007), e de acordo com López-Villavicencio e Mignon (2011), o crescimento do PIB é influenciado pelo nível de inflação (apesar da relação não ser linear), mostrando a forte relação entre os indicadores remanescentes.

Algumas dificuldades podem ser destacadas quando o objeto de estudo são indicadores econômicos. Uma delas é a disponibilidade e confiabilidade dos dados, como apontado por Hirschberg et al. (1991). Em um levantamento iniciado desde a década de 1960, um banco de dados com mais de 200 países e 22 indicadores, apenas 84 países possuíam dados mais recentes, considerando o intervalo de 1995 a 2014 em dez indicadores. Assim, partindo-se de uma matriz de dados de mais de 4000 observações que em potencial poderiam ser analisadas, na realidade somente 840 observações puderam ser incluídas. Tang e Salvador (1986) afirmam que a confiabilidade dos dados também diminui, principalmente em levantamentos de países pobres ou pouco desenvolvidos. Mas as técnicas de clusterização podem desempenhar bem mesmo nesse contexto (HIRSCHBERG et al., 1991).

Uma limitação da análise de agrupamento em oposição a outros tipos de análise (como regressão, por exemplo) é que não existe uma prática padrão para determinar onde o agrupamento deve ocorrer (WOLFSON et al., 2004). Os *clusters* são formados usando conhecimento de cunho exploratório do pesquisador, e não se apoiam em um padrão tido como verdadeiro na geração do modelo e dos agrupamentos.

Para futuros estudos, sugere-se que outras técnicas para validação da clusterização sejam aplicadas. Também para seleção de variáveis, no lugar de uma sistemática que omite variáveis, propõe-se o desenvolvimento de uma abordagem que atribua pesos para medir relevância de cada uma no agrupamento.

3.6 Referências

ABDI, H.; WILLIAMS, L. J. *Principal component analysis*. Wiley interdisciplinary reviews: computational statistics, v. 2, n. 4, p. 433-459, 2010.

ABIZADEH, S.; BASILEVSKY, A. *Socioeconomic classification of countries: a maximum likelihood factor analysis technique*. Social Science Research, v. 15, n. 2, p. 97-112, 1986.

ANDERTON, R.; TEWOLDE, T. *The global financial crisis: Understanding the global trade downturn and recovery1*. The World Economy, v. 34, n. 5, p. 741-763, 2011.

ANZANELLO, M. J.; FOGLIATTO, F. S. *Selecting the best clustering variables for grouping mass-customized products involving workers' learning*. International Journal of Production Economics, v. 130, n. 2, p. 268-276, 2011.

APERGIS, N.; CHRISTOU, C.; MILLER, S. *Convergence patterns in financial development: evidence from club convergence*. Empirical Economics, v. 43, n. 3, p. 1011-1040, 2012.

AUSLOOS, M; LAMBIOTTE, R. *Clusters or networks of economies? A macroeconomy study through Gross Domestic Product*. Physica A: Statistical Mechanics and its applications, v. 382, n. 1, p. 16-21, 2007.

BARRO, R. J. *Economic growth in a cross section of countries*. The quarterly journal of economics, v. 106, n. 2, p. 407-443, 1991.

BERLAGE, L.; TERWEDUWE, D. *The classification of countries by cluster and by factor analysis*. World Development, v. 16, n. 12, p. 1527-1545, 1988.

BRETERNITZ, V. J.; SILVA, L. A. *Big data: Um novo conceito gerando oportunidades e desafios*. Revista Eletrônica de Tecnologia e Cultura, v. 2, n. 2, 2013.

BRUSCO, M. J.; CRADIT, J. DENNIS. *A variable-selection heuristic for K-means clustering*. Psychometrika, v. 66, n. 2, p. 249-270, 2001.

Business Software Alliance. *What's the Big Deal With Data?* Executive Summary, 2015.

CAMACHO, M.; PEREZ-QUIROS, G.; SAIZ, L. *Do European business cycles look like one?* Journal of Economic Dynamics and Control, v. 32, n. 7, p. 2165-2190, 2008.

CERVO, V. L.; ANZANELLO, M. J. *Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento kernel*. **Prod.[online]. ahead of print**, p. 0-0, 2015.

CONSIGLIERE, I.; CARNEIRO, J. T.M.; TINTNER, G. *Um modelo econométrico aplicado à economia brasileira*. Revista Brasileira de Economia, v. 24, n. 1, p. 5-30, 1970.

DUDOIT, S.; FRIDLYAND, J. *A prediction-based resampling method for estimating the number of clusters in a dataset*. Genome biology, v. 3, n. 7, p. research0036. 1, 2002.

EINAV, L.; LEVIN, J. *The data revolution and economic analysis*. Innovation Policy and the Economy, v. 14, n. 1, p. 1-24, 2014.

GANDOMI, A.; HAIDER, M. *Beyond the hype: Big data concepts, methods, and analytics*. International Journal of Information Management, v. 35, n. 2, p. 137-144, 2015.

GIL, A. C. *Como classificar as pesquisas*. _____. Como elaborar projetos de pesquisa, v. 4, p. 41-56, 2002.

GUNVIG, A.; HANSEN, F.; BORGGAARD, C. *A mathematical model for predicting growth/no-growth of psychrotrophic C. botulinum in meat products with five variables*. Food control, v. 29, n. 2, p. 309-317, 2013.

GUYON, I.; ELISSEEFF, A. *An introduction to variable and feature selection*. Journal of machine learning research, v. 3, n. Mar, p. 1157-1182, 2003.

HAIR, J. F. et al. *Análise multivariada de dados*. Bookman Editora, 2005.

HIRSCHBERG, J. G.; MAASOUMI, E.; SLOTTJE, D. J. *Cluster analysis for measuring welfare and quality of life across countries. Journal of econometrics*, v. 50, n. 1-2, p. 131-150, 1991.

HUANG, J. Z. et al. *Automated variable weighting in k-means type clustering*. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 27, n. 5, p. 657-668, 2005.

JAIN, A. K. *Data clustering: 50 years beyond K-means*. Pattern recognition letters, v. 31, n. 8, p. 651-666, 2010.

JOHNSON, R.A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 3 ed. Prentice-Hall International (1992).

KHRAMOV, V.; LEE, J. R. *The Economic Performance Index (EPI): an Intuitive Indicator for Assessing a Country's Economic Performance Dynamics in an Historical Perspective*. International Monetary Fund, 2013.

KILSZTAJN, S. *Produto, capital e taxa de lucro—países industrializados e América Latina*. América Latina, v. 547, p. 933,779, 1998.

KUBRUSLY, L. S. *Um procedimento para calcular índices a partir de uma base de dados multivariados*. Pesqui. Oper., Rio de Janeiro , v. 21, n. 1, p. 107-117, jun. 2001 .

LIU, Y. et al. *Understanding of internal clustering validation measures*. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010. p. 911-916.

LLETI, R. et al. *Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes*. Analytica Chimica Acta, v. 515, n. 1, p. 87-100, 2004.

LÓPEZ-VILLAVICENCIO, A; MIGNON, V. *On the impact of inflation on output growth: Does the level of inflation matter?* Journal of Macroeconomics, v. 33, n. 3, p. 455-464, 2011.

MONTANARI, M. G.; MIRANDA, R. A. M.; GIRALDI, J. M. E. *Uma comparação quantitativa entre dois índices de competitividade*. Revista de Globalización, Competitividad y Gobernabilidad, v. 8, n. 2, p. 121, 2014.

PERRICONE, C. *Clustering Macroeconomic Variables*. 2013.

PINTO, L. W. *Aplicação de Modelos de Misturas Finitas na Classificação de Áreas de Risco para Tuberculose*. Tese (Doutorado em Engenharia Biomédica) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

RAFTERY, A. E.; DEAN, N. *Variable selection for model-based clustering*. Journal of the American Statistical Association, v. 101, n. 473, p. 168-178, 2006.

RENCHER, A. *Methods of Multivariate Analysis*, Wiley-Interscience (2002).

RESENDE, M. F. C.; VIEIRA, F. A. C. *Taxa de câmbio real e déficits gêmeos na economia monetária: relações de causalidade*. ECONOMIA & TECNOLOGIA, 2011.

ROUSSEEUW, P. J. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of computational and applied mathematics, v. 20, p. 53-65, 1987.

SILVA, J. G. S. et al. *Uma Experiência Didática com Base no Emprego do Matlab nos Cursos de Graduação da Faculdade de Engenharia da UERJ*. In: World Congress on Engineering and Technology Education-WCETE. 2004. p. 1-5.

SOTOCA, J. M. ; PLA, F. *Supervised feature selection by clustering using conditional mutual information-based distances*. Pattern Recognition, v. 43, n. 6, p. 2068-2081, 2010.

STARCZEWSKI, A.; KRZYŻAK, A. *Performance evaluation of the silhouette index*. In: International Conference on Artificial Intelligence and Soft Computing. Springer International Publishing, 2015. p. 49-58.

STEINLEY, D.; BRUSCO, M. J. A *New Variable Weighting and Selection Procedure for K-means Cluster Analysis*. Multivariate behavioral research, v. 43, n. 1, p. 77-108, 2007.

STEINLEY, D.; BRUSCO, M. J. *Selection of variables in cluster analysis: An empirical comparison of eight procedures*. Psychometrika, v. 73, n. 1, p. 125, 2008.

SUN, W. et al. *Regularized k-means clustering of high-dimensional data and its asymptotic consistency*. Electronic Journal of Statistics, v. 6, p. 148-167, 2012.

TANG, J.; SALVADOR, P. A. *Classification of countries for international development planning using cluster analysis*. Socio-Economic Planning Sciences, v. 20, n. 4, p. 237-241, 1986.

WANG, K.; WANG, B.; PENG, L. *CVAP: validation for cluster analyses*. Data Science Journal, v. 8, p. 88-93, 2009.

WITTEN, D. M.; TIBSHIRANI, R. *A framework for feature selection in clustering*. Journal of the American Statistical Association, v. 105, n. 490, p. 713-726, 2010.

WOLFSON, M.; MADJD-SADJADI, Z.; JAMES, P. *Identifying national types: A cluster analysis of politics, economics, and conflict*. Journal of Peace Research, v. 41, n. 5, p. 607-623, 2004.

YEDLA, M.; PATHAKOTA, S. R.; SRINIVASA, T. M. *Enhancing K-means clustering algorithm with improved initial center*. International Journal of computer science and information technologies, v. 1, n. 2, p. 121-125, 2010.

ZIMMER, J.; ANZANELLO, M. J. *Um novo método para seleção de variáveis preditivas com base em índices de importância*. Produção, São Paulo, v. 24, p. 84-93, 2014.

4 Terceiro Artigo: Análise da consistência de índices de avaliação de qualidade no agrupamento de indicadores econômicos de países

Paula Lunardi de Mello

Michel José Anzanello

Resumo

Em um mundo onde dados que cresce rapidamente e fornece informações em grande quantidade em tempo reduzido, estar munido de técnicas que transformem estes dados em informações de apoio à decisão se torna um diferencial. O desafio surge desde a coleta e a apuração dos dados até sua análise e visualização, e a ciência caminha na busca de desenvolver capacidades para descobrir as informações valiosas que existem em grandes bancos de dados. A análise de componentes principais (ACP) é das mais populares técnicas estatísticas multivariadas, e analisa uma base de dados descrita por variáveis independentes, objetivando extrair informações relevantes dos dados através de um conjunto de variáveis ortogonais chamadas componentes principais (ABDI & WILLIAMS, 2010). Este artigo aplica a ACP em indicadores econômicos de países trazidos pelo Banco Mundial, construindo na sequência agrupamentos através da técnica *k-means*. A qualidade dos agrupamentos gerados é avaliada através de três índices de avaliação, e os resultados são então discutidos.

Palavras-chave: Indicadores Econômicos, Análise de Componentes Principais, Índices de Avaliação da Qualidade de Clusterização.

4.1 Introdução

Em cenários onde dados são coletadas em volume e velocidade que crescem de maneira vertiginosa, estar munido de técnicas que transformem estes dados em informações de apoio à decisão se torna um diferencial. Segundo Einav & Levin (2014), ao longo das próximas décadas, o domínio de métodos capazes de tratar e entender essas grandes bases de dados para dar apoio à tomada de decisão mudará o panorama da política econômica

internacional. O desafio surge desde a coleta dos dados até sua manipulação, análise e interpretação, e a ciência caminha na busca de desenvolver capacidades para descobrir as informações valiosas que existem em grandes bancos de dados (CHEN & ZHANG, 2014).

Por outro lado, o custo do gerenciamento de um elevado volume de dados pode inviabilizar o seu tratamento, justificando a busca por modelos compostos por um número reduzido de variáveis, os quais podem significar uma substancial redução de custos e esforço computacional (ZIMMER & ANZANELLO, 2014).

Embora o grande número de variáveis disponíveis para análise pode tornar a interpretação de dados mais complicada e demorada, e uma técnica frequentemente usada para transpor esse obstáculo é a análise de componentes principais (ACP). Esta técnica combina variáveis potencialmente correlacionadas em componentes principais e, como resultado, traz uma simplificação para análise (OLAWALE & GARWE, 2010). A ACP é das mais populares e antigas técnicas estatísticas multivariadas, sendo usada por diversas áreas da ciência. A ACP analisa uma base de dados descrita por um elevado número de variáveis independentes, geralmente inter-correlacionadas, objetivando extrair informações relevantes dos dados através de um conjunto de variáveis ortogonais chamadas componentes principais (ABDI & WILLIAMS, 2010). Shlens (2014) afirma que um dos motivos da ampla aplicação da ACP se dá por revelar estruturas subjacentes simples em complexos conjuntos de dados.

Dentre as diversas áreas que podem ser beneficiadas com o processamento apropriado de dados está a economia. Identificar dados e informações relevantes passíveis de utilização pelos formuladores de políticas pode se tornar uma ferramenta para o desenvolvimento econômico dos países (TAYLOR & SCHROEDER, 2015). Dentre os benefícios de tais análises, Taylor e Schroeder (2015) afirmam que seria possível desenvolver um sistema sensível a mudanças nos padrões de indicadores econômicos, podendo gerar um alerta precoce de ação. Dessa maneira, indicadores descrevendo dimensões econômicas de países podem ser objeto de estudo de técnicas de análise multivariada de dados (dentre as quais a ACP) a fim de se identificar os indicadores (ou variáveis) que melhor caracterizam o nível de desenvolvimento de países. As evidências geradas pelas análises podem auxiliar na elaboração de políticas de crescimento apropriadas, tendo-se as variáveis principais identificadas como norteadoras para diferentes grupos de países.

A literatura aponta que a redução da dimensão da matriz de dados como um passo anterior ao processo de agrupamento pode trazer benefícios em termos de interpretação e precisão dos grupos gerados. Isso porque o subespaço de dados obtidos pela redução de variáveis pode ser uma forma de tornar a formação dos *clusters* mais natural e intuitiva (DING & LI, 2007). Para comprovar se esta formação é mais robusta, métricas de avaliação de clusterização podem ser aplicadas.

Nesse sentido, este artigo aplica a ACP em indicadores econômicos disponibilizados pelo Banco Mundial, os quais descrevem diversas dimensões econômicas de um elevado número de países. Os componentes principais gerados pela ACP são então utilizados como variáveis de clusterização com vistas à construção de agrupamentos de países com perfis similares de desenvolvimento econômico através da técnica *k-means*. Ao final, avalia-se a qualidade dos diferentes agrupamentos formados utilizando três diferentes métricas para quantificação da consistência dos grupos gerados.

O trabalho está dividido em cinco seções. A seção 4.2 traz a revisão da literatura para análise de componentes principais, método *k-means* e índices para avaliação da qualidade de agrupamentos. A seção 4.3 descreve o método para a avaliação da qualidade dos arranjos, aplicando três diferentes índices. Na seção 4.4 os resultados encontrados são analisados e, finalmente, na seção 4.5 são colocadas as conclusões e discussões geradas pelo estudo.

4.2 Revisão da literatura

A revisão da literatura deste artigo consiste em explorar aplicações da análise de componentes principais, passando para o emprego do método *k-means* na clusterização de dados, e finalizando com métricas de avaliação da qualidade de agrupamentos.

4.2.1 Análise de componentes principais (ACP)

A ACP é uma técnica estatística multivariada usada para reduzir o número de variáveis em um conjunto de dados criando um novo conjunto de menor dimensão (VYAS &KUMARANAYAKE, 2006). Para Hosseini e Kaneko (2001) a ACP reduz o número de variáveis para um menor número de componentes que representam a maior parte da variação das variáveis originais. Em termos matemáticos, a partir de um conjunto inicial X de dados com *n* variáveis correlacionadas, a ACP gera *m* componentes ou índices não-correlacionados,

onde cada componente é uma combinação linear ponderada das variáveis iniciais (VYAS &KUMARANAYAKE, 2006; BRAGA et al., 2012). Então, em conjuntos de variáveis x_1 até x_n,

$$CP_1 = a_{11}x_1 + a_{12}x_2 + ...+a_{1n}x_n$$

$$\vdots$$

$$CPm = a_{m1}x_1 + a_{m2}x_2 + ...+a_{mn}x_n$$

onde a_{mn} representa o peso da m-ésima componente principal da n-ésima variável.

Os pesos de cada componente principal são dados pelos autovetores da matriz de correlação da matriz de covariância, se os dados forem padronizados (VYAS & KUMARANAYAKE, 2006; ABDI & WILLIAMS, 2010; BRAGA et al., 2012; SHLENS, 2014). Braga et al. (2012) descrevem de maneira algébrica como são obtidas as componentes principais. A partir de uma matriz de dados normalizados **X**, composta por n observações e p características, é possível obter a matriz de correlação **R** através da Eq. (4.1):

$$\mathbf{R} = \frac{1}{n}\mathbf{X}.\mathbf{X}^t \tag{4.1}$$

$\mathbf{X}^t$ é a transposta de **X**.

Sendo **R** uma matriz simétrica e positiva de dimensão (p x p), sua diagonalização é feita por uma matriz **A** de mudança de base, chamada de matriz de autovetores. A matriz diagonal **D**, cujos elementos diagonais são os autovalores de **R**, é dada pela Eq. (4.2):

$$\mathbf{D} = \mathbf{A}^{-1}.\mathbf{R}.\mathbf{A} \tag{4.2}$$

onde $\mathbf{A}^{-1}$ é a inversa da matriz **A**.

Por combinações lineares entre a transposta dos autovetores de $\mathbf{A}^t$ e a matriz de observações **X**, são obtidas as componentes principais U_1, U_2, ..., U_p, ou seja (4.3):

$$\mathbf{U} = \mathbf{A}^t.\mathbf{X} \tag{4.3}$$

Os coeficientes a_{jk} representam os pesos associados a cada componente U_k e são ordenados por ordem decrescente dos autovalores mais significativos de **A**.

A distribuição espacial das componentes principais representa sua contribuição em relação à variância total dos dados originais. Os fatores comuns são obtidos pela relação entre a matriz dos autovetores (**A**) e a raiz quadrada do autovalor (λ), ou seja, o primeiro fator comum é obtido pela relação entre o primeiro autovetor (a_{11}) e o primeiro autovalor (λ_1) e assim de maneira sucessiva (BRAGA et al., 2012). Rencher (2002) afirma que na análise de componentes principais se busca maximizar a variância de uma combinação linear das variáveis. A primeira componente principal é a combinação linear com variância máxima, e se procura uma dimensão ao longo da qual as observações são separadas ou espalhadas ao máximo (RENCHER, 2002). A segunda componente principal é a combinação linear com variância máxima em uma direção ortogonal à primeira, e assim sucessivamente, como mostrado na Figura 3.1 (RENCHER, 2002; VYAS & KUMARANAYAKE, 2006; BRAGA et al., 2012).

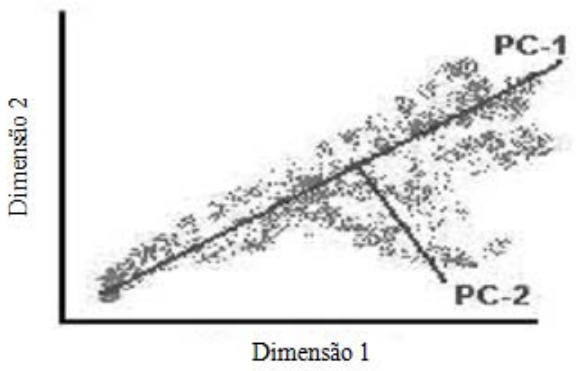

Figura 4.1: Representação da ortogonalidade das componentes principais (VYAS & KUMARANAYAKE, 2006).

Diversas aplicações da ACP podem ser encontradas na literatura. Favero et al. (2005) aplicaram a técnica em dados de indicadores macroeconômicos dos Estados Unidos e países da zona do Euro para realizar análise empírica da política monetária desses países. Kolenikov e Angeles (2009) aplicaram ACP a variáveis econômicas ordinais e binárias, comprovando que o desempenho da técnica depende do nível de correlação entre as variáveis. O uso da ACP ordinal é recomendado se as categorias definidas são confiáveis e bem estabelecidas. Olawale e Garwe (2010) usaram a ACP para identificar os maiores obstáculos para o crescimento de novas micro e pequenas empresas na África do Sul. Divididas entre obstáculos

internos e externos, as 30 variáveis iniciais foram reduzidas a cinco, reduzindo a complexidade da análise.

Outra aplicação foi realizada por Vyas e Kumaranayake (2006), com o objetivo de desenvolver um índice de status socioeconômico, sem partir de dados de renda ou consumo, e realizando ACP sobre indicadores de ativos duráveis, acesso a infraestrutura e habitação. Os resultados do estudo foram tidos como válidos e robustos (VYAS & KUMARANAYAKE, 2006).

Hassani et al. (2012) realizaram a análise de componentes principais em múltiplos blocos, ou de múltiplas amostras, para comparar os parâmetros globais dos dados e os encontrados nos blocos. Em dados de bioespectroscopia e outros simulados, mostraram que existe dificuldade em interpretar os resultados dos blocos em conexão com os dados globais. Por sua vez, Hosseini e Kaneko (2011) utilizaram indicadores de mais de cem países para aplicar a ACP encontrar os que eram determinantes para a sustentabilidade. Concluíram que quando os pilares institucional, ambiental e econômico estão fortemente correlacionados, o desenvolvimento econômico não pode explicar a deterioração ambiental.

Por fim, na pesquisa de Braga et al. (2012), a ACP foi usada para estimar séries temporais de precipitação pluviométricas no Estado do Rio Grande do Norte. Partindo de uma clusterização *k-means* para identificar as regiões do estado, os pesquisadores utilizaram dados de duas regiões para obter os estimadores das componentes principais, encontrando resultados com erro aceitável e definindo o método como eficaz.

4.2.2 Método de clusterização *k-means*

Apesar do método *k-means* ter sido proposto há mais de 50 anos e diversos outros algoritmos de clusterização terem sido publicados desde então, ele ainda é amplamente utilizado (JAIN, 2010) devido à sua aplicação rápida e eficaz (YEDLA et al., 2010). Ghosh e Dubey (2013) afirmam que clusterização é basicamente um método de particionamento aplicado para analisar dados e tratar observações dos dados como objetos baseados em locais e distâncias. Dividir os objetos em *clusters* mutuamente exclusivos (*k*) se realiza de tal maneira que objetos de um mesmo *cluster* permanecem o mais próximo possível uns dos outros, mas o mais longe possível de objetos de outros *clusters* (GHOSH & DUBEY, 2013;

RENCHER, 2002). O algoritmo *k-means* encontra o número desejado de *clusters* distintos e seus centroides (GHOSH & DUBEY, 2013). Um centroide é o ponto cujas coordenadas são obtidas por meio de cálculo da média de dada uma das coordenadas dos objetos atribuídos aos *clusters* (GHOSH & DUBEY, 2013).

De forma complementar, Johnson e Wichern (1992) afirmam que o *k-means* parte de um número inicial de *k* grupos e agrupa cada observação ao centroide (valor médio das observações contidas no agrupamento) mais próximo. Este método tipicamente utiliza a distância Euclidiana para identificar diferentes *clusters* nos dados, embora outras possam ser utilizadas (SUN et al., 2012; LEE, 2013).

Aplicações do método são muitas em diversas áreas. Kanungo et al. (2002) aplicaram a clusterização *k-means* em bancos de dados de quantização de cores e segmentação de imagens, enquanto que Oyelade et al. (2010) aplicaram a técnica para agrupar alunos de acordo com seu desempenho acadêmico.

A integração de ACP e técnicas de clusterização também tem encontrado espaço na literatura. Din e He (2004) analisaram dados de expressão gênica e grupos de notícia da internet aplicando ACP e *k-means*, enquanto que Rodrigues (2010) identificou os componentes principais de medição de impulsos de resposta a estímulos visuais em corujas e com *k-means* buscou agrupar neurônios de acordo com sua sensibilidade aos impulsos.

É conhecido que geralmente uma estrutura de agrupamento de interesse ocorre em um subespaço definido por um subconjunto das variáveis inicialmente selecionadas, sendo importante revelá-lo para encontrar a real estrutura de agrupamento (HUANG et al., 2005; RAFTERY & DEAN, 2006). Sun et al. (2012) colocam que, apesar do método *k-means* ser simples e computacionalmente eficiente, seu desempenho pode ser deteriorado significativamente na presença de variáveis irrelevantes. Muitas delas podem não conter informações sobre a estrutura de agrupamento e até tornar a análise muito complexa (SUN et al., 2012). Assim, identificar as variáveis com maior poder de separação dos dados é uma etapa fundamental para transpor essas dificuldades no agrupamento de dados (JAIN, 2010; WITTEN & TIBSHIRANI, 2010; SUN et al., 2012).

4.2.3 Índices para avaliação da qualidade de agrupamentos

Um dos desafios fundamentais da clusterização consiste em apropriadamente avaliar os resultados sem informações auxiliares, dada a característica exploratória da técnica (RENDÓN et al., 2011). Halkidi et al. (2001) complementam que, como a clusterização é um método não supervisionado (o qual não indica *a priori* o número real de *clusters* presentes em um conjunto de dados), faz-se necessário validar os resultados.

Tal validação é reconhecida como etapa essencial para o sucesso da aplicação da clusterização (LIU et al., 2010), e um dos cursos de ação consiste na utilização de índices de avaliação da qualidade (RENDÓN et al., 2011). Handl et al. (2005) afirmam que compreender que o uso de técnicas de validação analíticas por si só não é suficiente, mas que a compreensão dos princípios de funcionamento dos algoritmos de agrupamento, medidas de validação e suas interações é crucial para permitir a validação objetiva do *cluster*.

Em termos gerais, existem duas abordagens para avaliar a qualidade de agrupamentos: índices internos e índices externos (HALKIDI et al., 2001; DUDOIT et al., 2002; HANDL et al., 2005; WANG et al., 2009; RENDÓN et al., 2011). Os índices internos avaliam o resultado com respeito a uma informação intrínseca aos dados isoladamente (RENDÓN et al., 2011). A solução é avaliada usando parâmetros inerentes ao conjunto de dados, em termos quantitativos que envolvem vetores do próprio conjunto de dados (HALKIDI et al., 2001). Dentre os índices internos, medidas como compacidade e separabilidade são usadas (HALKIDI et al., 2001; DUDOIT et al., 2002): (i) a compacidade mede com base na variância ou distância o quão proximamente estão relacionadas às observações de um *cluster*, enquanto a (ii) separabilidade mede o quão distinto ou bem separado um *cluster* é em relação a outros *clusters* (HALKIDI et al., 2001; LIU et al., 2010). Os índices internos mais utilizados incluem s*ilhouette index*, índice Davies-Bouldin, índice Calinski-Harabasz, índice Dunn e o índice *Root-mean-square standard deviation* (WANG et al., 2009); tais índices são agora detalhados.

O s*ilhouette index* (SI) é um índice composto que reflete a compacidade e a separação de *clusters*. Esse índice varia de -1 a +1 e o número ótimo de *clusters* é determinado pela maximização desse índice, que mede o quão distintos estão uns dos outros os *clusters* formados; um SI médio maior indica uma melhor qualidade geral do resultado de agrupamento (WANG et al., 2009). O SI baseia-se nas diferenças entre as distâncias das

observações, analisadas aos pares dentro do *cluster* (LIU et al., 2010). Conforme Anzanello e Fogliatto (2011), o SI_n, referente à observação *n*, é calculado como na Eq. (4.4):

$$\mathrm{SI}_n = \frac{b(n) - a(n)}{\max\{b(n), a(n)\}} \tag{4.4}$$

sendo *a(n)* a média das distâncias da *n*-ésima observação dentro do mesmo *cluster*, e *b(n)* a média das distâncias da mesma observação em relação as outras alocadas no *cluster* mais próximo.

O SI não depende da técnica de clusterização utilizada na análise, uma vez que é calculado com base em distâncias. Normalmente é usado para medir a qualidade global da clusterização através da média de SI_n, conforme Eq. (4.5):

$$\overline{\mathrm{SI}} = \frac{\sum_{n=1}^{N} \mathrm{SI}_n}{N} \tag{4.5}$$

onde *n* é o índice da observação e *N* o total de observações.

O índice Davies-Bouldin (DB) é uma medida da semelhança média entre cada grupo e seu mais semelhante, onde valores baixos correspondem a *clusters* que são compactos e têm centros que estão distantes um dos outros (WANG et al., 2009). Para Rendón et al. (2011), o DB visa identificar agrupamentos que são compactos e separados adequadamente. Ao minimizar este índice, os *clusters* são os mais distintos entre si, e consequentemente se consegue a melhor partição (LIU et al., 2010). Segundo Rendón et al. (2011) é calculado conforme a Eq. (4.6):

$$DB = \frac{1}{c}\sum_{i=1}^{c} Max_{i \neq j}\left\{\frac{d(X_i) + d(X_j)}{d(c_i, c_j)}\right\} \tag{4.6}$$

onde c é o número de *clusters*; i, j são identificadores dos *clusters*, então $d(X_i)$ e $d(X_j)$ são as distâncias de todas as observações nos *clusters* i e j para seus respectivos centroides, d (c_i, c_j) é a distância entre esses centroides.

O índice Calinski-Harabasz (CH) mede o isolamento entre *clusters* e a coerência dentro dos *clusters* (WANG et al., 2009). Seu valor máximo determina o número recomendado de *clusters* (DUDOIT et al., 2002). Para Caliński e Harabasz (1974) seu cálculo é dado pela Eq. (4.7):

$$CH = \frac{SS_b}{SS_w} \times \frac{N-k}{k-1} \tag{4.7}$$

onde SS_b é a variância total entre *clusters*, SS_w é a variância total dentro dos *clusters*, k é o número de *clusters* e N é o número total de observações. A variância total entre *clusters* SS_b é dada por (4.8):

$$SS_b = \sum_{i=1}^{k} n_i \, \|m_i - m\|^2 \tag{4.8}$$

onde k é o número de *clusters*, m_i é o centroide do *cluster i*, m é a média de todas observações, e $\|m_i - m\|$ é a distância euclidiana entre os dois vetores. A variância total dentro dos *clusters* SS_w é igual (4.9):

$$SS_w = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2 \tag{4.9}$$

onde k é o número de *clusters*, x é uma dada observação, c_i é o i-ésimo *cluster*, m_i é o centroide do *cluster i*, e $\|x - m_i\|$ é a distância euclidiana entre os dois vetores. *Clusters* bem definidos têm alta variância total entre *clusters* (SS_b) e baixa variância total dentro dos *clusters* (SS_w). Quanto maior o CH, melhor é o agrupamento.

Apesar dos diversos índices disponíveis para avaliação da qualidade dos agrupamentos gerados, para Handl et al. (2005), devido ao viés intrínseco a muitas técnicas de validação interna, uma análise cuidadosa dos resultados obtidos é necessária, e os resultados devem ser sempre verificados com técnicas alternativas de validação complementar.

A segunda abordagem para avaliação da qualidade de agrupamentos se dá com índices externos. Rendón et al. (2011) definem os índices externos como aqueles que avaliam o resultado em relação a uma estrutura pré-especificada. Esses índices comparam a estrutura de grupos descoberta com uma estrutura de grupos previamente conhecida (DUDOIT et al., 2002), refletindo uma intuição sobre a estrutura de agrupamento do conjunto de dados (HALKIDI et al., 2001). As medidas de validação externa reconhecem o número real de *clusters* em avançado, e são usadas principalmente para escolher um algoritmo de agrupamento em um conjunto de dados específicos (LIU et al., 2010). Na prática, informações externas muitas vezes não estão disponíveis, portanto, nesta situação, medidas de validação interna são a única opção para validação da solução. Exemplos de índices externos podem ser encontrados em Desgraupes (2013).

4.3 Método para avaliação de agrupamento de países

O objetivo da clusterização é encontrar uma estrutura em um conjunto de dados e, portanto, é de natureza exploratória. De tal forma, seu objetivo é exploratório, pois busca aprofundar os conhecimentos no tema (GIL, 2002). Esta pesquisa pode ainda ser classificada como aplicada, utilizando abordagem quantitativa, uma vez que analisa dados reais de indicadores de desempenho da economia de países.

O estudo utilizou dados do Banco Mundial reunidos pelo *Global Federation of Competitiveness Councils* (GFCC), que é uma rede mundial de líderes de conselhos de competitividade. Ela define direcionadores (*drivers*) fundamentais para a competitividade nacional e que sustentam o crescimento econômico global. Para realizar o presente estudo, foram usados dez indicadores econômicos agrupados pertencentes ao *driver* denominado "desempenho geral", apresentados no Anexo A. O presente estudo restringiu-se a países que possuíam os referidos indicadores econômicos no intervalo de 1995 a 2014 (visto que muitos países passaram a ter seus dados coletados depois de 1995).

Na sequência, foi calculada a mediana dos países para cada indicador (84 países e 10 indicadores), por ser uma estatística menos suscetível a valores discrepantes. Foi então realizada a análise de componentes principais, buscando reduzir o número de indicadores (variáveis). A decomposição dos autovalores foi feita a partir da matriz de covariância dos dados. Foram priorizadas as componentes que explicavam pelo menos 90% da variância dos dados. Todos os procedimentos computacionais foram realizados no software Matlab.

Os componentes principais retidos (representados por U) são então utilizados como variáveis de entrada na clusterização, sendo os agrupamentos formados através do *k-means*. A utilização dos componentes principais como variáveis de clusterização (ao invés das variáveis originais) visa a reduzir o efeito da correlação entre as variáveis sobre a clusterização. O procedimento de clusterização é rodado para um intervalo de número de agrupamentos (variando-se o valor de k), e a qualidade de cada solução avaliada através dos índices *silhouette index* (SI), Calinski-Harabasz (CH) e Davies-Bouldin (DB). Neste estudo, variou-se o número de *clusters* de 2 a 42. Por fim, identificou-se o número (k) recomendado de *clusters* objetivando otimizar cada um dos índices.

4.4 Resultados

A partir do banco de dados dos países e respectivos indicadores (Anexo B), foi aplicada a ACP e encontrados os pesos das componentes principais para cada indicador (*a*), como mostra Tabela 4.1.

Tabela 4.1: Pesos das componentes principais para cada variável (indicador)

Componente Principal (CP_m)	Pesos da CP_m									
	a_{m1}	a_{m2}	a_{m3}	a_{m4}	a_{m5}	a_{m6}	a_{m7}	a_{m8}	a_{m9}	a_{m10}
CP_1	0,0498	-0,3166	0,5032	0,3321	-0,2877	-0,0654	-0,1873	-0,3678	-0,4878	-0,1961
CP_2	0,434	-0,3353	0,1104	-0,2298	0,5084	0,098	-0,0821	0,4565	-0,3852	-0,0523
CP_3	0,5562	0,1949	-0,2387	0,2554	-0,2125	-0,0089	-0,0535	0,1112	0,1863	-0,6598
CP_4	-0,1154	0,111	0,0921	0,133	-0,035	0,9104	-0,3285	0,0763	0,0647	0,0104
CP_5	-0,0005	-0,0076	0,0824	0,0097	-0,0285	0,3362	0,9065	-0,0602	-0,1716	-0,1562
CP_6	-0,0112	-0,2657	0,4017	0,3073	-0,2937	-0,0989	0,1333	0,6043	0,4001	0,1856
CP_7	0,372	-0,3318	0,2125	-0,0653	0,285	0,1024	0,0417	-0,5136	0,5797	0,094
CP_8	0,5536	0,2258	-0,1647	0,2361	-0,2144	0,0517	0,0586	-0,0714	-0,2194	0,674
CP_9	-0,161	0,158	-0,0342	0,7445	0,6169	-0,101	0,0514	-0,012	-0,0104	0,005
CP_{10}	0,1248	0,693	0,6538	-0,2127	0,1326	-0,1048	-0,0075	0,0205	0,0309	-0,0374

Na análise dos pesos das CP's, vê-se que os maiores valores absolutos (módulos de a_{mn}) se distribuem de maneira uniforme entre as variáveis originais (indicadores). Pode-se observar que CP_2 e CP_4 são menos impactadas pelo indicador *População total* (I10), enquanto para CP_1 isso ocorre para o indicador *Exportação de bens e serviços* (I1) e para CP_3 o indicador *Importação de bens e serviços* (I6) recebe menor peso . Quanto ao I1, verificando resultados de diversos estudos de indicadores econômicos, seu impacto na avaliação econômica de países não recebe destaque (BARRO, 1991; ANDERTON & TEWOLDE, 2011; KHRAMOV & LEE, 2013). O menor peso do I10 para CP_2 e CP_4 pode ser explicada porque impacta indiretamente na economia. PIB, inflação e outras variáveis podem medir de maneira mais efetiva e direta o desempenho econômico dos países (CONSIGLIERE & CARNEIRO, 1970).

Como resultado da ACP também foram encontradas as variâncias das componentes principais, que correspondem aos autovalores da matriz de covariância dos dados originais. Os valores são mostrados na Tabela 4.2. Enfatiza-se o elevado nível de variância explicado pelas primeiras quatro componentes, os quais respondem por 90% da variabilidade existente nos dados. De tal forma, as quatro primeiras CP's são utilizados na etapa seguinte do método.

Tabela 4.2: Variância das componentes principais, em valor e percentual

CP_m	**Variância**	**Variância relativa**	**Variância acumulada**
CP_1	0,0275	46%	46%
CP_2	0,0145	24%	70%
CP_3	0,0075	12%	82%
CP_4	0,0048	8%	90%
CP_5	0,0037	6%	96%
CP_6	0,0013	2%	98%
CP_7	0,0006	1%	99%
CP_8	0,0002	0%	100%
CP_9	0,0001	0%	100%
CP_{10}	0,0001	0%	100%

Para encontrar o número recomendado de *clusters* para o agrupamento (*k*), foram calculados os índices internos SI, CH e DB para o intervalo de 2 a 42 *clusters*. A evolução do valor de cada um dos índices para o intervalor de *k* estudado pode ser observada na Figura 4.2, enquanto que os resultados recomendados são mostrados na Tabela 4.3.

Tabela 4.3: Valores de índices de avaliação de agrupamento e respectivo *k*

Índice	**Valor ótimo**	***k* recomendado**
SI	0,7654	3
CH	143,2652	28
DB	0,5098	40

Por fim, o índice SI atinge seus melhores valores no início do intervalo, e apresenta redução substancial a partir de $k = 6$. Já o CH vai alcançando maiores valores com o aumento

de *k*, mantendo um patamar de desempenho a partir de *k* = 20. Inversamente, o índice DB tem característica menor-melhor, e vai ganhando performance no incremento de *k*.

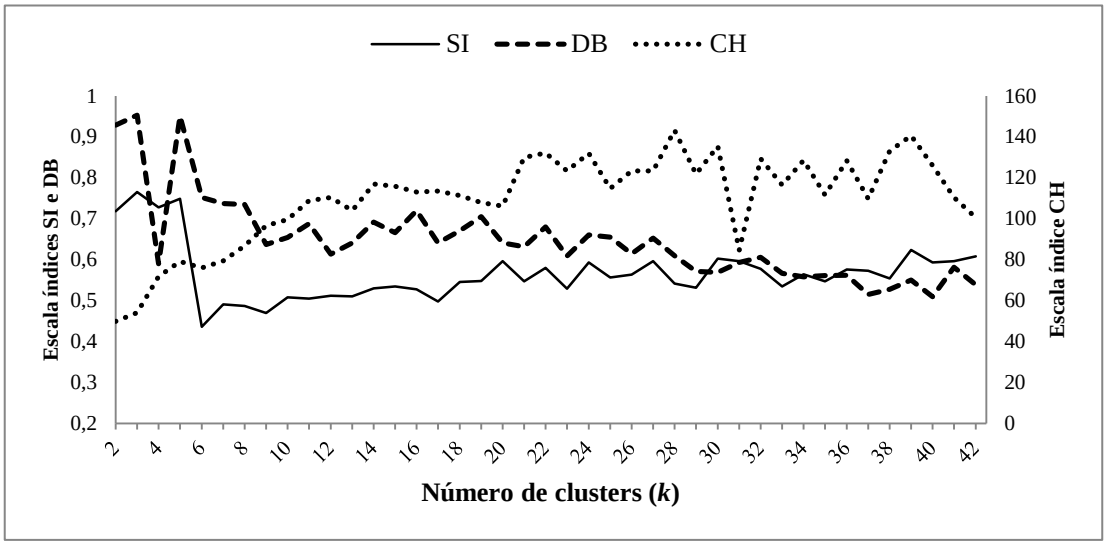

Figura 4.2: Evolução dos índices internos para o intervalo de *k*

O SI busca maximizar a compacidade e a separabilidade dos *clusters* (RENDÓN et al., 2011), sendo *k* recomendado para chegar ao valor máximo de SI igual a três. Por medir a distância média entre as observações, é menos sensível a *outliers*, o que pode gerar uma configuração mais compacta de número de *clusters*.

O valor ótimo do índice CH foi obtido com 28 *clusters*. Este índice objetiva encontrar o agrupamento que tenha alta variação entre *clusters* e uma baixa variação dentro do *cluster*. Então isto pode explicar um número relativamente alto de *k* quando relacionado ao número de países (84), podendo as observações dentro do grupo, nessa configuração, serem menos distintas entre si e distintas entre grupos.

O índice DB mede a similaridade média entre os *clusters* em relação ao seu mais semelhante. Para uma clusterização adequada é necessário que sejam compactos e separados, ou seja, quanto menor o DB melhor a solução de agrupamento (KOVÁCS, 2005). O melhor índice foi encontrado na solução com 40 *clusters*. Um maior número de *clusters* pode gerar *clusters* mais compactos e menos discrepantes entre si, já que as observações seriam mais próximas no *cluster* e menos similares entre *clusters* (HALKIDI et al., 2001).

4.5 Discussões e conclusões

Este artigo utilizou a ACP em indicadores econômicos de países para gerar agrupamentos através da técnica *k-means*, objetivando otimizar o resultado de três diferentes índices: SI, CH e DB.

Bro e Smilde (2014) afirmam que a ACP é uma ferramenta poderosa e versátil, pois é capaz de revelar as relações entre variáveis e amostras com base em um reduzido número de combinações das variáveis originais. Inicialmente, a matriz de dados avaliada continha 84 observações e dez indicadores e foi reduzida a quatro componentes principais, tornando a análise mais objetiva e sucinta. No que tange o banco de dados de indicadores de desempenho de países, a quantidade de variáveis disponível pode ser muito alta, ou, por outro lado, inexistente ou não-confiável. Por isso encontrar formas de contornar esses obstáculos e trazer informações relevantes a respeito da estrutura dos dados se mostra importante. Processar esse imenso banco de maneira efetiva, confiável e rápida é fundamental para que estratégias sejam definidas e adotadas em tempo apropriado, sendo que ACP se mostra uma ferramenta adequada nesse sentido.

Quanto às métricas para avaliação da qualidade da clusterização, diferentes tipos de índices são usados para resolver diferentes questões de pesquisa (RENDÓN et al., 2011). A seleção do índice depende do tipo de informação disponível e do escopo do estudo a ser realizado. Exemplos disso foram as diferentes soluções encontradas para o arranjo de países nos três índices avaliados. Enquanto DB e CH encontram melhores valores em maiores *k*, o SI encontrou seu valor ótimo em um arranjo de três *clusters*. Dessa forma, se a pesquisa visa encontrar agrupamentos mais coesos e discrepantes entre si, os índices DB e CH podem ser aplicados, o que vai exigir um estudo mais aprofundado dos diversos grupos gerados, entrando no âmbito local da economia dos países.

Já no caso de se buscar arranjos mais amplos, ou globais no caso de indicadores econômicos, o SI pode ser a escolha para identificar agrupamentos com tendências e padrões mais gerais de desempenho, voltando a análise para a macroeconomia de blocos.

Para futuros estudos, podem ser explorados índices externos de avaliação da qualidade, já que muitos indicadores econômicos possuem padrão e comportamento conhecido e podem ser fonte de informação para a formação de agrupamentos.

4.6 Referências

ABDI, H.; WILLIAMS, L. J. *Principal component analysis*. Wiley interdisciplinary reviews: computational statistics, v. 2, n. 4, p. 433-459, 2010.

ANDERTON, R.; TEWOLDE, T. *The global financial crisis: Understanding the global trade downturn and recovery1*. The World Economy, v. 34, n. 5, p. 741-763, 2011.

BARRO, R. J. *Economic growth in a cross section of countries*. The quarterly journal of economics, v. 106, n. 2, p. 407-443, 1991.

BEVERIDGE, S.; NELSON, C. R. *A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'*. Journal of Monetary economics, v. 7, n. 2, p. 151-174, 1981.

BRAGA, C. C. et al. *Aplicação dos Componentes Principais na Simulação e Consistência de Séries Temporais*. Revista de Geografia (UFPE), v. 29, n. 1, p. 113-125, 2012.

BRO, R.; SMILDE, A. K. *Principal component analysis*. Analytical Methods, v. 6, n. 9, p. 2812-2831, 2014.

CALIŃSKI, T.; HARABASZ, J. *A dendrite method for cluster analysis*. Communications in Statistics-theory and Methods, v. 3, n. 1, p. 1-27, 1974.

CHEN, C. P.; ZHANG, C. *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*. Information Sciences, v. 275, p. 314-347, 2014.

CONSIGLIERE, I.; CARNEIRO, J. T.M.; TINTNER, G. *Um modelo econométrico aplicado à economia brasileira*. Revista Brasileira de Economia, v. 24, n. 1, p. 5-30, 1970.

DESGRAUPES, B. *Clustering indices*. University of Paris Ouest-Lab Modal'X, v. 1, p. 34, 2013.

DING, C.; HE, X. *K-means clustering via principal component analysis*. In: Proceedings of the twenty-first international conference on Machine learning. ACM, 2004. p. 29.

DING, C.; LI, T. *Adaptive dimension reduction using discriminant analysis and k-means clustering*. In: Proceedings of the 24th international conference on Machine learning. ACM, 2007. p. 521-528.

DUDOIT, S.; FRIDLYAND, J. *A prediction-based resampling method for estimating the number of clusters in a dataset*. Genome biology, v. 3, n. 7, p. research0036. 1, 2002.

EINAV, L.; LEVIN, J. *The data revolution and economic analysis*. Innovation Policy and the Economy, v. 14, n. 1, p. 1-24, 2014.

FAVERO, C. A.; MARCELLINO, M.; NEGLIA, F. *Principal components at work: the empirical analysis of monetary policy with large data sets*. Journal of Applied Econometrics, v. 20, n. 5, p. 603-620, 2005.

GHOSH, S.; DUBEY, S. K. *Comparative analysis of k-means and fuzzy c-means algorithms*. International Journal of Advanced Computer Science and Applications, v. 4, n. 4, 2013.

GIL, A. C. *Como classificar as pesquisas*. _____. Como elaborar projetos de pesquisa, v. 4, p. 41-56, 2002.

HALKIDI, M.; BATISTAKIS, Y.; Vazirgiannis, M. *On clustering validation techniques*. Journal of intelligent information systems, v. 17, n. 2-3, p. 107-145, 2001.

HANDL, J.; KNOWLES, J.; KELL, D. B. *Computational cluster validation in post-genomic data analysis*. Bioinformatics, v. 21, n. 15, p. 3201-3212, 2005.

HOSSEINI, H. M.; KANEKO, S. *Dynamic sustainability assessment of countries at the macro level: A principal component analysis*. Ecological indicators, v. 11, n. 3, p. 811-823, 2011.

JAIN, A. K. *Data clustering: 50 years beyond K-means*. Pattern recognition letters, v. 31, n. 8, p. 651-666, 2010.

JOHNSON, R.A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 3 ed. Prentice-Hall International (1992).

KANUNGO, T. et al. *An efficient k-means clustering algorithm: Analysis and implementation*. IEEE transactions on pattern analysis and machine intelligence, v. 24, n. 7, p. 881-892, 2002.

KHRAMOV, V.; LEE, J. R. *The Economic Performance Index (EPI): an Intuitive Indicator for Assessing a Country's Economic Performance Dynamics in an Historical Perspective*. International Monetary Fund, 2013.

KOLENIKOV, S.; ANGELES, G. *Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?*. Review of Income and Wealth, v. 55, n. 1, p. 128-165, 2009.

KOU, G.; PENG, Y.; WANG, G. *Evaluation of clustering algorithms for financial risk analysis using MCDM methods*. Information Sciences, v. 275, p. 1-12, 2014.

KOVÁCS, F.; LEGÁNY, C.; BABOS, A. *Cluster validity measurement techniques*. In: 6th International symposium of hungarian researchers on computational intelligence. 2005.

LIU, Y. et al. *Understanding of internal clustering validation measures*. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010. p. 911-916.

MARTINEZ, W. L. et al. *Exploratory data analysis with MATLAB*. CRC Press, 2010.

OLAWALE, F.; GARWE, D. *Obstacles to the growth of new SMEs in South Africa: A principal component analysis approach*. African journal of Business management, v. 4, n. 5, p. 729, 2010.

OYELADE, O. J.; OLADIPUPO, O. O.; OBAGBUWA, I. C. *Application of k Means Clustering algorithm for prediction of Students Academic Performance*. arXiv preprint arXiv:1002.2425, 2010.

RENCHER, A. *Methods of Multivariate Analysis*, Wiley-Interscience (2002).

RENDÓN, E. et al. *Internal versus external cluster validation indexes*. International Journal of computers and communications, v. 5, n. 1, p. 27-34, 2011.

RODRIGUES, B. D. *Análise em componentes principais e k-means aplicados a estudos de sensibilidade de células neuronais a estímulos*. Proceedings Seminário Interno da disciplina de Redes Neurais ArtificiaiS, p. 67, 2010.

RIBEIRO, I. S. *Sobre as diferenças na aplicação de SVD e PCA: Um estudo pragmático*. 2013.

SHLENS, J. *A tutorial on principal component analysis*. arXiv preprint arXiv:1404.1100, 2014.

SILVA, J. G. S. et al. *Uma Experiência Didática com Base no Emprego do Matlab nos Cursos de Graduação da Faculdade de Engenharia da UERJ*. In: World Congress on Engineering and Technology Education-WCETE. 2004. p. 1-5.

TAYLOR, L.; SCHROEDER, R. *Is bigger better? The emergence of big data as a tool for international development policy*. GeoJournal, v. 80, n. 4, p. 503-518, 2015.

VYAS, S.; KUMARANAYAKE, L. *Constructing socio-economic status indices: how to use principal components analysis.* Health policy and planning, v. 21, n. 6, p. 459-468, 2006.

WANG, K.; WANG, B.; PENG, L. *CVAP: validation for cluster analyses*. Data Science Journal, v. 8, p. 88-93, 2009.

ZIMMER, J.; ANZANELLO, M. J. *Um novo método para seleção de variáveis preditivas com base em índices de importância*. Produção, São Paulo, v. 24, p. 84-93, 2014.

5 Considerações Finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

O objetivo principal da dissertação foi aplicar sistemáticas de agrupamentos de países com base em indicadores econômicos (variáveis de clusterização). A proposta se mostrou válida, pois as séries temporais dos indicadores são extensas e numerosas em variáveis. Além disso, a abordagem proposta vem para auxiliar na redução da dimensão da matriz de dados e tornar a analise menos complexa e mais focada nos indicadores relevantes para os agrupamentos. A aplicação técnica *k-means* se mostrou adequada para agrupamento dos países a partir de indicadores econômicos. Os *clusters* formados mostraram distinção entre si, permitindo que os agrupamentos gerados pudessem ser avaliados. A análise dividida por décadas também possibilitou observar a evolução dos indicadores, identificando o comportamento dos países ao longo dos dados temporais.

Os objetivos específicos foram definidos como: *i*) selecionar os indicadores mais relevantes para clusterização de países pela técnica "*leave one variable out at a time"*; *ii)* aplicar a técnica de redução dimensional ACP; e *iii*) avaliar a qualidade de *clusters* gerados por meio da aplicação dos índices internos SI, CH e DB.

O objetivo (*i*) foi atingido no segundo artigo, no qual os indicadores retidos pela técnica "*leave one variable a time*" resultaram em SI = 0,9628, indicando agrupamentos de boa qualidade. *PIB*, crescimento *PIB* e *Inflação* foram os indicadores mais relevantes e reforçam o exposto na literatura, a qual define esses três indicadores como os que melhor representam o desempenho econômico de países (juntamente com a taxa de desemprego e a taxa de déficit publico), os quais compõem a equação do índice de desempenho econômico.

O terceiro artigo vai ao encontro do objetivo (*ii*), aplicando a ACP para reduzir o número de indicadores econômicos e focar a análise nos mais significativos, tornando a análise mais objetiva e sucinta. A redução da matriz inicial de dados foi de 60% no número de observações.

O objetivo (*iii*) foi alcançado pelos três artigos, os dois primeiros aplicando o SI e o último agregando o cálculo dos índices CH e DB. Foi possível detectar as diferentes

abordagens de cada um dos índices, uma vez que cada índice é usado para resolver diferentes questões. DB e CH têm seus valores otimizados em maiores *k*, enquanto que o SI encontrou seu valor ótimo em $k = 3$. Caso o estudo vise identificar grupos coesos e distintos entre si, os índices DB e CH podem conduzir a melhores resultados, mas irão exigir análise mais focada na economia local dos países. O SI pode ser a escolha caso se busque agrupamentos mais abrangentes, sendo possível identificar tendências e padrões mais gerais de desempenho, voltando a análise para a macroeconomia de blocos.

A validação dos agrupamentos formados também se mostrou etapa importante, uma vez que ela destaca diferentes características dos grupos, dependendo do índice aplicado. Por isso a escolha da métrica deve ter como critério o objetivo do estudo, assim como mostrou a análise de cada um dos índices no último artigo.

5.2 Sugestões para trabalhos futuros

Desdobramentos dessa dissertação podem ser desenvolvidos, tais como:

a) Substituir os dados de mediana das décadas pelos valores anuais no procedimento de clusterização;

b) Integrar análises sociais e políticas aos resultados de agrupamentos de países;

c) Aplicar outras técnicas multivariadas de seleção de variáveis para indicadores econômicos com vistas à clusterização e classificação; e

d) Avaliar a qualidade de agrupamentos de países através de índices externos.

ANEXO A - Indicadores de desempenho geral geral da competitividade dos países

#	**Indicador**	**Medida**	**Definição**
I1	Exportação de bens e serviços	Percentual do Produto Interno Bruto (PIB)	Exportações de bens e serviços representam o valor de todos os bens e outros serviços de mercado fornecidos ao resto do mundo. Eles incluem o valor das mercadorias, fretes, seguros, transportes, viagens, royalties, taxas de licença, e outros serviços, tais como comunicação, construção, financeiro, informação, negócio, pessoal, e os serviços governamentais. Exclui a compensação de funcionários e rendimentos de investimentos (anteriormente chamado de fator de serviços) e pagamentos de transferências.
I2	Produto Interno Bruto (PIB) per capita	Moeda, em milhares de dólares	PIB per capita é o produto interno bruto dividido pela população na metade do ano. O PIB é a soma do valor agregado bruto por todos os produtores residentes na economia mais quaisquer impostos sobre os produtos e menos quaisquer subsídios não incluídos no valor dos produtos; é calculado sem fazer deduções para depreciação de ativos fabricados ou para o esgotamento e a degradação dos recursos naturais
I3	Produto Interno Bruto	Moeda, em bilhões de dólares	O PIB é a soma do valor agregado bruto por todos os produtores residentes na economia mais quaisquer impostos sobre os produtos e menos quaisquer subsídios não incluídos no valor dos produtos; é calculado sem fazer deduções para depreciação de ativos fabricados ou para o esgotamento e a degradação dos recursos naturais
I4	Crescimento do PIB	Percentual	Taxa anual de crescimento percentual do PIB a preços de mercado com base em moeda local constante
I5	Formação Bruta de Capital Fixo	Percentual do PIB	A formação bruta de capital fixo inclui melhoramentos fundiários , compras de máquinas, equipamentos, a construção de estradas, ferrovias e similares, incluindo escolas, escritórios, hospitais, residências particulares e edifícios comerciais e industriais
I6	Importação de bens e serviços	Percentual do PIB	As importações de bens e serviços representam o valor de todos os bens e outros serviços de mercado recebidos do resto do mundo. Incluem o valor da mercadoria, frete, seguro, transporte, viagens, royalties, licenças e outros serviços, tais como serviços de comunicação, construção, financeiro, informações, negócios, pessoais e

#	Indicador	Medida	Definição
			governamentais
I7	Rendimento Nacional Bruto (RNB) per capita com base na paridade do poder de compra (PPC)	Moeda, em milhares de dólares	O RNB PPC é o rendimento nacional bruto convertido em dólares internacionais usando as taxas de paridade do poder de compra. Um dólar internacional tem o mesmo poder de compra sobre o RNB do que um dólar americano nos Estados Unidos.
I8	Formação Bruta de Capital Fixo Nominal	Moeda, em bilhões de dólares	A formação bruta de capital fixo inclui melhoramentos fundiários , compras de máquinas, equipamentos, a construção de estradas, ferrovias e similares, incluindo escolas, escritórios, hospitais, residências particulares e edifícios comerciais e industriais
I9	Inflação dos preços ao consumidor	Percentual	Inflação medida pelo índice de preços ao consumidor que reflete a variação percentual anual no custo para o consumidor médio de aquisição de uma cesta de bens e serviços que possam ser fixados ou alterados em intervalos específicos, tais como anual
I10	População total	Número, em milhões	A população total conta todos os residentes independentemente do estatuto legal ou da cidadania - com exceção dos refugiados que não estão permanentemente estabelecidos no país de asilo, que são geralmente considerados parte da população do seu país de origem

Fonte: Banco Mundial

ANEXO B – Tabela de medianas dos indicadores por país (período de 1995 a 2014)

País	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
África do Sul	28,264	5,168	243,183	3,070	17,821	27,038	9,395	41,041	5,780	47,038
Albânia	21,908	2,624	7,921	5,600	24,891	46,521	6,225	1,878	2,652	3,019
Alemanha	36,641	34,385	2836,550	1,650	20,108	31,666	31,675	580,597	1,526	82,074
Argélia	37,217	2,851	94,262	3,500	25,314	25,250	10,125	21,790	4,094	33,043
Armênia	23,566	1,404	4,239	6,900	22,044	47,700	4,515	1,157	4,564	3,020
Austrália	19,594	32,223	653,105	3,715	26,670	20,912	30,890	175,269	2,641	20,261
Azerbaijão	46,006	1,312	10,963	9,655	25,204	38,569	5,690	5,240	3,223	8,349
Bahamas	43,996	22,002	7,400	2,140	25,740	48,870	21,570	1,751	1,795	0,326
Bangladesh	13,890	0,474	67,276	5,266	25,411	19,358	1,730	17,105	6,878	141,887
Bélgica	71,614	36,238	378,700	1,756	22,310	68,624	32,825	82,758	1,807	10,450
Benim	20,503	0,522	4,204	4,524	19,479	29,883	1,380	0,815	3,090	8,053
Bielorússia	60,167	2,752	26,676	5,672	26,422	66,276	9,095	6,935	35,468	9,697
Botsuana	50,321	4,997	9,444	5,939	27,929	43,514	9,015	2,505	8,030	1,850
Brasil	11,738	5,221	889,222	3,110	18,599	12,108	10,200	165,186	6,466	187,298
Bulgária	49,053	3,558	27,610	3,162	21,019	52,467	9,590	6,527	6,078	7,761
Burkina Faso	11,070	0,389	5,151	6,487	21,975	26,474	1,045	1,008	2,326	13,225
Burundi	7,486	0,153	1,059	3,903	16,151	30,361	0,595	0,172	10,248	7,798
Butão	38,009	1,183	0,761	6,891	46,726	57,479	3,840	0,412	6,645	0,643
Camarões	21,429	0,904	16,182	4,095	17,739	20,991	2,165	2,906	2,887	17,897
Camboja	54,752	0,440	5,815	7,191	17,675	63,017	1,565	1,084	3,961	13,216
Canadá	36,593	33,929	1091,273	2,680	21,531	33,055	34,045	235,312	1,954	32,154
Chade	33,863	0,557	5,531	5,677	22,373	41,016	1,245	1,349	4,725	9,889
Chipre	56,798	24,158	17,717	3,398	21,126	58,082	25,025	3,079	2,443	1,024
Cingapura	191,717	28,637	120,802	6,151	26,947	170,571	49,390	35,287	1,372	4,221
Colômbia	16,289	3,073	131,821	4,008	21,601	19,937	7,705	25,435	6,127	43,005
Coreia do Sul	37,731	17,130	831,509	4,713	30,707	34,420	23,475	257,907	2,859	48,089
Costa do Marfim	45,082	0,936	16,820	1,891	11,861	37,768	2,305	1,770	2,525	17,968
Costa Rica	42,333	4,571	19,281	4,389	19,492	45,177	9,120	3,601	10,519	4,217
Dinamarca	46,156	47,652	257,901	1,389	20,812	40,358	33,820	54,013	2,105	5,412
Dominica	35,324	5,290	0,373	1,283	18,907	53,003	7,050	0,072	1,405	0,071
Egito	20,659	1,406	98,735	4,557	17,930	25,780	7,075	18,889	7,416	74,269
Equador	26,998	2,865	39,049	3,909	20,556	27,703	7,070	7,843	6,544	13,622
Etiópia	12,502	0,153	11,266	9,369	29,142	28,749	0,615	3,632	8,108	75,558
Filipinas	44,013	1,178	97,221	4,874	20,498	48,828	5,155	20,310	4,738	85,369
França	26,704	34,377	2163,895	1,960	21,556	26,653	30,185	465,360	1,692	62,942
Gabão	58,837	6,311	8,608	3,551	25,066	30,041	13,400	1,797	1,674	1,363
Geórgia	30,473	1,328	5,768	5,666	25,138	48,558	4,025	1,528	6,372	4,435
Granada	29,007	6,295	0,647	2,166	32,691	54,274	9,090	0,164	2,203	0,103
Grécia	21,257	21,680	238,621	3,014	22,249	30,614	25,015	35,452	3,352	11,023
Guatemala	24,793	1,963	25,588	3,653	16,386	36,212	5,340	4,687	6,587	13,031
Guiné-Bissau	18,710	0,383	0,559	2,983	6,365	32,174	1,170	0,051	3,315	1,447
Holanda	65,490	40,439	659,208	1,968	21,223	58,190	36,450	134,357	2,065	16,301
Honduras	51,080	1,352	9,222	3,795	24,406	66,080	3,235	2,393	7,906	6,817

País	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
Hong Kong	188,940	26,990	179,461	3,064	23,595	180,189	35,175	44,427	2,223	6,798
Índia	18,416	0,685	777,900	7,483	28,634	20,666	2,700	230,124	6,767	1135,373
Indonésia	29,622	1,207	271,353	5,266	26,564	26,416	5,595	67,444	6,500	224,762
Islândia	35,320	40,980	13,042	3,662	21,388	40,072	32,465	2,225	3,995	0,294
Israel	34,235	20,640	137,621	4,032	19,511	35,641	24,355	26,673	2,404	6,870
Itália	24,955	31,581	1826,319	1,380	19,953	24,609	29,215	366,049	2,143	57,827
Jordânia	47,570	2,241	11,998	4,202	25,304	71,770	8,130	3,330	3,227	5,351
Macedônia	37,552	2,926	5,971	3,012	21,374	53,263	7,620	1,206	2,283	2,041
Malásia	102,834	5,244	134,142	5,717	23,017	90,291	14,750	40,788	2,384	25,564
Malawi	28,007	0,233	2,710	4,600	14,187	41,895	0,530	0,478	15,077	12,578
Marrocos	29,782	1,900	58,236	4,012	26,904	36,120	4,545	15,676	1,386	30,239
Maurícia	58,261	5,173	6,335	3,917	23,259	65,065	11,115	1,364	5,709	1,225
Mauritânia	43,499	0,649	2,009	5,225	27,465	54,250	2,595	0,824	5,044	3,109
México	26,648	7,388	818,307	3,090	21,068	28,001	11,450	173,549	4,860	109,003
Moçambique	27,968	0,295	6,138	7,342	18,493	38,463	0,625	1,241	9,689	20,825
Moldávia	48,937	0,776	2,793	4,699	22,610	78,728	3,155	0,643	11,695	3,600
Mongólia	51,539	0,898	2,258	6,691	28,952	61,260	5,080	0,630	9,201	2,512
Nepal	15,142	0,304	7,702	4,473	21,211	33,272	1,465	1,550	7,537	25,352
Níger	17,705	0,244	3,229	3,960	18,710	28,708	0,670	0,609	2,464	13,241
Noruega	40,476	62,173	286,540	2,315	21,639	28,604	46,085	57,315	2,211	4,608
Paquistão	14,397	0,683	103,740	3,960	15,471	19,388	3,515	16,898	7,805	151,812
Paraguai	53,269	1,952	9,877	4,150	15,934	46,422	4,655	1,863	7,698	5,750
Peru	22,339	2,580	70,821	5,614	21,395	19,855	6,050	12,826	3,298	27,442
Portugal	27,373	18,415	193,246	1,682	23,183	37,126	20,865	34,921	2,730	10,458
Quênia	22,297	0,503	17,417	4,277	18,602	32,559	1,950	3,060	9,306	34,893
Reino Unido	27,081	37,693	2303,519	2,553	18,177	28,379	34,495	379,519	2,108	60,195
República Dominicana	31,255	3,314	30,001	5,722	19,956	38,068	7,040	5,311	6,400	9,170
Romênia	32,948	4,092	87,484	3,754	22,468	41,097	8,875	20,010	10,433	21,386
Santa Lúcia	47,305	5,536	0,910	0,944	26,126	63,289	8,330	0,233	3,137	0,164
São Cristóvão e Nevis	36,591	10,701	0,522	3,715	42,264	52,805	16,555	0,219	2,309	0,049
São Vicente e Granadinas	35,613	4,936	0,536	1,752	25,900	57,648	7,650	0,138	1,797	0,109
Senegal	27,078	0,752	8,369	3,800	22,833	40,007	1,795	1,909	1,563	11,118
Sudão	14,403	0,606	23,991	5,958	19,530	16,891	2,095	5,637	13,776	31,583
Suécia	44,037	42,764	385,374	2,741	22,118	38,460	34,630	84,261	0,963	9,012
Tailândia	69,524	2,937	179,150	4,677	26,094	61,611	8,900	60,908	2,887	65,634
Tanzânia	18,107	0,399	14,878	5,782	23,654	26,172	1,445	3,550	7,561	38,500
Togo	37,330	0,368	2,026	3,753	15,898	53,419	1,040	0,319	2,429	5,503
Turquia	23,156	6,486	437,576	5,715	20,334	26,473	10,750	90,674	10,514	67,417
Ucrânia	48,438	1,598	75,512	2,500	19,674	51,408	6,185	16,541	12,074	47,278
Uganda	13,770	0,308	8,477	6,640	20,804	25,132	1,050	1,792	6,344	27,579
Uruguai	24,670	7,129	23,690	4,381	16,465	25,450	10,720	3,749	8,106	3,327