

**UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E
AMBIENTAL**

**CIÊNCIA DOS DADOS E INTELIGÊNCIA ARTIFICIAL
PARA AGRUPAMENTO E DIAGNÓSTICO DE CIDADES
INTELIGENTES**

ALEXANDRE NISGOSKI DA PAIXÃO

ORIENTADOR: FRANCISCO EVANGELISTA JUNIOR

**MONOGRAFIA DE PROJETO FINAL EM ENGENHARIA
CIVIL 2**

BRASÍLIA/DF: OUTUBRO/2021

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

**CIÊNCIA DOS DADOS E INTELIGÊNCIA ARTIFICIAL PARA
AGRUPAMENTO E DIAGNÓSTICO DE CIDADES INTELIGENTES**

ALEXANDRE NISGOSKI DA PAIXÃO

MONOGRAFIA DE PROJETO FINAL 2 SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM ENGENHARIA CIVIL.

APROVADA POR:

Francisco Evangelista Junior, PhD (ENC-UnB)
(ORIENTADOR)

Marcos Honorato de Oliveira, DSc. (ENC-UnB)
(EXAMINADOR INTERNO)

Yaeko Yamashita, PhD (FAU-UnB)
(EXAMINADORA EXTERNA)

BRASÍLIA - DF, 29 DE OUTUBRO DE 2021.

FICHA CATALOGRÁFICA

NISGOSKI DA PAIXÃO, ALEXANDRE
CIÊNCIA DOS DADOS E INTELIGÊNCIA ARTIFICIAL PARA
AGRUPAMENTO E DIAGNÓSTICO DE CIDADES INTELIGENTES.

ix, 85 p., 297 mm (ENC/FT/UnB, Bacharel, Engenharia Ambiental, 2021)

Monografia de Projeto Final – Universidade de Brasília, Faculdade de Tecnologia.

Departamento de Engenharia Civil e Ambiental.

1. APRENDIZADO ESTATÍSTICO 2. CIDADES INTELIGENTES

3. INDICADORES

4. NORMA ISO

I. ENC/FT/UnB

II. Bacharel

REFERÊNCIA BIBLIOGRÁFICA

PAIXÃO, A. N. (2021). *CIÊNCIA DOS DADOS E INTELIGÊNCIA ARTIFICIAL PARA AGRUPAMENTO E DIAGNÓSTICO DE CIDADES INTELIGENTES*. Monografia de Projeto Final, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, p. 60.

CESSÃO DE DIREITOS

NOME DO AUTOR: Alexandre Nisgoski da Paixão

TÍTULO DA MONOGRAFIA DE PROJETO FINAL: Ciência dos dados e inteligência artificial para agrupamento e diagnóstico de cidades inteligentes.

GRAU / ANO: Bacharel em Engenharia Civil / 2021

É concedida à Universidade de Brasília a permissão para reproduzir cópias desta monografia de Projeto Final e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta monografia de Projeto Final pode ser reproduzida sem a autorização por escrito do autor.

ALEXANDRE NISGOSKI DA PAIXÃO

alexnisgoski@gmail.com

AGRADECIMENTOS

Agradeço à minha família por todo o apoio, logístico e emocional, ao longo desta jornada. A mascote gatinha de minha irmã, Carlota Joaquina, que também se mostrou muito disposta a ouvir minhas ideias e receber meus carinhos. Minha gata provou-se ainda mais resiliente, como aliás tem escrito no braço, por aturar-me ao longo deste período. Não menos importantes foram meus amigos de curso, os BIIIIIRLs, já há muito amigos do peito e de copo. Meus amigos da Liga também em muito colaboraram com histórias e experiências que me mantiveram no rumo, em meio a tantos que o perderam. Além destes, claro, todos aqueles espalhados pelo Brasil, em especial na minha segunda casa, João Pessoa, onde ficam os integrantes do meu amado Donzelão. Impossível seria deixar de agradecer também ao meu orientador, que tanto me mostrou dentro da Engenharia, alumando-me os caminhos dentro desta ciência melindrosa.

Por fim, gostaria de ressaltar a importância da bandeira brasileira – e desta ideia difusa de grandeza que a rodeia -, em todos os momentos pendurada em minha parede enquanto escrevia esta monografia, sempre me lembrando que, embora estejamos vivendo dias sombrios e tortuosos, dias cansativos e por vezes desanimadores, com toda a certeza neste mundo, o nosso futuro espelhará esta grandeza.

RESUMO

Cidades inteligentes são tão difíceis de serem classificadas que nem mesmo têm uma definição própria e aceita por todos. Foram estudadas diversas normas dos últimos 20 anos que vieram trazendo novas e velhas maneiras de se estabelecer um conjunto de indicadores que permitissem o diagnóstico, o acompanhamento e, claro, a classificação das cidades pelo quesito da inteligência. Em especial, uma publicação recente da ISO (Organização Internacional de Padronização), a ISO 37122, editada em português pela ABNT (Associação Brasileira de Normas Técnicas), que trouxe diversos indicadores novos.

Métodos de aprendizado estatístico vêm sendo utilizados há muitos anos com intuito de permitir a visualização intuitiva de dados muito complexos para serem trabalhados em estado bruto. A separação de algumas técnicas deste tipo, aptas ao caso estudado, foi realizada neste trabalho. Além da visualização facilitada, fez-se uso destas técnicas para realizar o agrupamento das cidades com base nos valores apresentados dos indicadores.

Em um primeiro momento, um teste-piloto foi realizado com dados fictícios criados pelo autor, validando as técnicas apresentadas. Então, buscaram-se dados reais para testagem em campo dos métodos escolhidos. Partindo-se ao uso em dados reais, mostraram-se as diferenças que existiam entre os dados reais e os dados sintéticos, as possíveis diferenças de análise e os agrupamentos reais. Além disso, foi apresentada uma metodologia de redução do número de indicadores necessários para geração dos mesmos agrupamentos com base nas correlações entre os indicadores, permitindo a economia computacional e financeira de estudos reais de grande vulto.

Concluiu-se que as técnicas de baixa dimensionalidade (duas dimensões) não são tão eficazes para dados não tão dicotômicos e mais nuançados. As técnicas tridimensionais, no entanto, mostram-se eficazes para a geração de agrupamentos. A redução da dimensionalidade do problema com base na matriz de correlação mostrou-se eficaz em manter os mesmos agrupamentos, mas não a com base na PCA. No total, atingiu-se uma redução de 27,8% no número de indicadores para formar os mesmos agrupamentos. Foram indicados parâmetros a serem seguidos no tocante à matriz de correlação de variáveis. Novos estudos e testes foram sugeridos para complementação do tema.

Palavras-chave: Cidades inteligentes. Indicadores. ISO 37122. ISO 37120. NBR. Agrupamentos. Aprendizado estatístico. Biplot. Dendrograma. Escalonamento multidimensional. Análise de componentes principais.

ABSTRACT

Smart cities are so hard to classify that they don't even have an agreed upon definition. Many standards from the last twenty years were studied, they have been bringing new and old ways to establish a set of indicators that would allow the diagnosis, follow-up and, of course, classification of cities based on how smart they are. In particular, a recent publication by ISO, ISO 37122, edited in Portuguese by ABNT, that brought several new indicators.

Statistical learning methods have been under use for several years with the purpose to allow for the intuitive visualization of a too complex dataset to be worked on in its brute state. A selection of some techniques of this kind, fit for the studied case, was made in this work. Beyond facilitated visualization, the use of these techniques was made to also define clusters of cities based on their indicators' values.

At first, a pilot-test was made with fictional data created by the author, in order to validate the techniques here presented. Then, real data was looked for and submitted to the same method. From this dataset of real values, it was demonstrated that there were differences that existed between real and synthetic data, and showed all possible new different analysis and the real clusters. Furthermore, a methodology for the reduction of the number of indicators necessary from the appropriate clustering was presented. This allows for computational and financial savings, especially in real large studies.

It was concluded that low dimensionality techniques (bidimensional) aren't as effective for not so dichotomous a dataset. Tridimensional techniques, however, have showed to be effective for the creation of clusters. Dimensionality reduction of the problem based on the correlation matrix has showed to be effective in keeping the same clusters. PCA however was not able to steadily form clusters with fewer indicators being fed into it. Some parameters to be followed in relation to the correlation matrix of variables were indicated. New studies and tests were also suggested for complementation of the topic covered.

Keywords: Smart Cities. Indicators. ISO 37122. ISO 37120. NBR. Clusters. Statistical learning. Biplot. Dendrogram. Multidimensional scaling. Principal component analysis.

SUMÁRIO

1. INTRODUÇÃO	1
2. OBJETIVOS	4
2.1. OBJETIVO GERAL.....	4
2.2. OBJETIVOS ESPECÍFICOS	4
3. FUNDAMENTAÇÃO TEÓRICA E REVISÃO	
BIBLIOGRÁFICA	5
3.1. INDICADORES PARA CIDADES INTELIGENTES E SUSTENTÁVEIS	6
3.1.1. O conceito de indicador	6
3.1.2. Conjuntos de indicadores e metodologias.....	8
3.1.2.1. Indicadores de Desenvolvimento Sustentável (2002).....	9
3.1.2.2. European Smart Cities Ranking (2007)	10
3.1.2.3. ISO 37120: Sustainable Development of Communities – Indicators for City Services and Quality of Life (2014).....	11
3.1.2.4. ETSI TS 103 463: Key Performance Indicators for Sustainable Digital Multiservice Cities (2017)	12
3.1.2.5. ISO 37122: Sustainable Cities and Communities – Indicators for Smart Cities (2019).....	13
3.2. AGRUPAMENTOS ESTATÍSTICOS.....	15
3.2.1. Dendrograma.....	15
3.2.2. Escalonamento Multidimensional (MDS)	17
3.2.3. Análise de Componente Principal (PCA)	20
4. METODOLOGIA E ANÁLISES INICIAIS	25
4.1. APRESENTAÇÃO E DEFINIÇÕES	25
4.2. ANÁLISES ESTATÍSTICAS	30
4.2.1. Dendrograma.....	30
4.2.2. Escalonamento multidimensional (MDS).....	31
4.2.3. Análise de componentes principais (PCA)	32
5. ESTUDO DE CASO: APLICAÇÃO A DADOS REAIS.....	35
5.1. COLETA DOS DADOS.....	35
5.2. ANÁLISES ESTATÍSTICAS SEM REDUÇÃO DE INDICADORES	38

5.2.1. Dendrograma.....	39
5.2.2. Escalonamento Multidimensional Bidimensional	39
5.2.3. Escalonamento Multidimensional Tridimensional	40
5.2.4. Análise de Componente Principal Tridimensional	42
5.3. ANÁLISES ESTATÍSTICAS COM REDUÇÃO DE INDICADORES	45
5.3.1. 13 Indicadores (5 a menos que o original).....	45
5.3.2. 12 Indicadores (6 a menos que o original).....	48
5.3.3. 11 Indicadores (7 a menos que o original).....	50
6. CONCLUSÕES	54
7. REFERÊNCIAS BIBLIOGRÁFICAS	57

LISTA DE FIGURAS

Figura 3.1 Taxonomia proposta por Huovila, Bosch e Airaksinen (2019).....	9
Figura 3.2 Exemplo de dendrograma (MATLAB)	16
Figura 3.3 Escalonamento Multidimensional clássico aplicado sobre dados de votos de membros do Congresso dos Estados Unidos (ANDREWMAN327, 2013)	20
Figura 3.4 Exemplo de <i>Scree plot</i> de uma Análise de Componente Principal (NGO, 2018)	23
Figura 3.5 Exemplo de <i>biplot</i> PCA (NGO, 2018)	24
Figura 4.1 Matriz de correlação e histogramas de indicadores para as cidades D, I, J, K e L	28
Figura 4.2 Distribuição dos indicadores nas cidades fictícias	29
Figura 4.3 Dendrograma com dados fictícios de cidades	30
Figura 4.4 Escalonamento Multidimensional com grupos gerados.....	31
Figura 4.5 <i>Scree Plot</i> para dados de indicadores de cidades inteligentes.....	32
Figura 4.6 <i>Biplot</i> de Análise de Componente Principal 3D com grupos gerados	33
Figura 4.7 <i>Biplot</i> de Análise de Componente Principal 2D com grupos gerados	34
Figura 5.1 Dendrograma, 18 indicadores	39
Figura 5.2 MDS 2D ou projeção XY, 18 indicadores	40
Figura 5.3 MDS 3D e projeção XZ, 18 indicadores.....	41
Figura 5.4 <i>Scree Plot</i> acumulado para 18 indicadores	43
Figura 5.5 <i>Biplot</i> 3D com 18 indicadores.....	43
Figura 5.6 <i>Scree Plot</i> acumulado para 13 indicadores	47
Figura 5.7 <i>Biplot</i> 3D com 13 indicadores.....	47
Figura 5.8 <i>Scree Plot</i> acumulado para 12 indicadores	49

Figura 5.9 <i>Biplot</i> 3D com 12 indicadores.....	50
Figura 5.10 <i>Scree Plot</i> acumulado para 12 indicadores	51
Figura 5.11 <i>Biplot</i> 3D com 11 indicadores.....	52

LISTA DE TABELAS

Tabela 3.1 Comparação entre normativos de indicadores para cidades inteligentes e sustentáveis em suas revisões mais recentes	14
Tabela 4.1 Conjunto de dados de cidades fictícias e seus indicadores simulados.....	25
Tabela 4.2 Indicadores ilustrativos selecionados	26
Tabela 4.3 Matriz de correlação entre cidades fictícias.....	27
Tabela 4.4 Grupos identificados no dendrograma.....	30
Tabela 4.5 Grupos identificados no escalonamento multidimensional	32
Tabela 4.6 Grupos identificados no PCA em 3D e em 2D.....	34
Tabela 5.1 Indicadores selecionados	36
Tabela 5.2 Cidades selecionadas	38
Tabela 5.3 Grupos formados a partir do MDS 3D	42
Tabela 5.4 Grupos formados a partir do <i>biplot</i> tridimensional com 18 indicadores	44
Tabela 5.5 Matriz de correlação entre indicadores reais (Parte 1)	45
Tabela 5.6 Matriz de correlação entre indicadores reais (Parte 2)	46
Tabela 5.7 Grupos formados a partir do <i>biplot</i> tridimensional com 13 indicadores	48

LISTA DE EQUAÇÕES

Equação 3.1	16
Equação 3.2	17
Equação 3.3	18
Equação 3.4	19
Equação 3.5	19
Equação 3.6	22
Equação 3.7	22
Equação 3.8	22
Equação 3.9	22

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
CODEPLAN	Companhia de Planejamento do Distrito Federal
Covid-19	Doença do Coronavírus 2019
ETSI	Instituto de Padronização em Telecomunicação Europeu
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IDS	Indicadores de Desenvolvimento Sustentável
IoT	Internet das Coisas
ISO	Organização Internacional de Padronização
nMDS	Escalonamento Multidimensional Não-métrico
NBR	Norma Brasileira
MDS	Escalonamento Multidimensional
mMDS	Escalonamento Multidimensional Métrico
ONU	Organização das Nações Unidas
PCA	Análise de Componente Principal
PNE	Portador de Necessidade Especial
PIB	Produto Interno Bruto
SARS-COV 2	Coronavírus 2 da Síndrome Respiratória Aguda Grave
SCMM	Modelo de Maturidade de Cidades Inteligentes
WCCD	Conselho Mundial de Dados de Cidade

1. INTRODUÇÃO

A população mundial ainda não parou de aumentar, mas a população rural sim (UNFPA, 2016). Isso significa que aumentos na população total do planeta já implicam em urbanização, ou seja, aumento da população nas cidades. Desde 2008, e pela primeira vez na História, mais da metade da população mundial é citadina (UNFPA, 2007). No Brasil, esta barreira foi rompida ainda na década de 70, com 51,5% dos brasileiros já vivendo em cidades no ano de 1980 (FAUSTO, 2019).

Esta mudança de eixo do campo para as cidades ocorre pois estas últimas têm sido o coração do desenvolvimento humano e avanço tecnológico, o que se reflete no geral em melhores oportunidades de trabalho, experiência cultural mais vasta e melhores condições de moradia (RIFFAT, POWELL e AYDIN, 2016). A urbanização pode ser positiva, muito embora a atual concentração de riquezas e aumento da desigualdade, crescimento de favelas, perturbação social e ambiental - para manter-se em eufemismo - em grande parte do mundo não teçam um cenário muito animador. Há de se lembrar, entretanto, que embora concentrem a pobreza, as cidades também representam a melhor esperança de erradicá-la (UNFPA, 2007).

Apesar de algumas cidades no passado terem sido planejadas, uma vasta parcela delas experimentou crescimento caótico, com provisão de serviços e projetos estratégicos quase não existentes, especialmente quando este processo foi movido por migração do campo durante períodos de industrialização e mecanização da agricultura, como no caso das cidades industriais britânicas no século XIX ou o Rio de Janeiro no século XX (RIFFAT, POWELL e AYDIN, 2016). Uma cena histórica brasileira que exemplifica os desdobramentos do caos do crescimento urbano foi a Revolta da Vacina, quando a população se insurgiu contra a vacinação da varíola, doença que pôde circular mais facilmente no agregado urbano desorganizado carioca.

A cena do caos causado pela varíola seria revisitada nos anos seguintes ao fim da Primeira Guerra Mundial com a gripe espanhola. A epidemia de 1918 evidenciou diversas falhas do aparato urbano da capital brasileira à época, especialmente das condições sanitárias e de moradia (GOULART, 2005). A gripe espanhola aleitou 66% da população da cidade durante toda a epidemia. Mais de cem anos depois, na pandemia da Covid-19 causada pelo vírus SARS-COV 2, as favelas cariocas, epítomes do crescimento urbano turbulento brasileiro, acumulam mais mortos que 162 países como Dinamarca, Austrália e Paraguai

(THUSWOHL, 2021), mostrando que o desenvolvimento urbano no Brasil ainda está muito inferior ao desejado e há muito espaço para progresso.

Em face às profundas disfuncionalidades nos modelos de como as cidades estão organizadas e como lidam com suas demandas, o debate acerca de novas maneiras de encará-las se acalorou. Uma das expressões que passou a integrar o léxico usado pela comunidade científica especializada no final da década de 90 foi “*smart city*” (ANTHOPOULOS, 2015). Uma terminologia razoavelmente amorfa e inespecífica até hoje, cidades inteligentes (como traduzido para o português) só despontaram e ficaram difundidas como conceito em 2010 (BERNADINI, VITERBO, *et al.*, 2017). Cidades inteligentes atualmente aparecem como novo paradigma de desenvolvimento urbano inteligente e crescimento socioeconômico sustentável (NEIROTTI, DE MARCO, *et al.*, 2014), implicando em modelos para municípios interconectados, previsíveis, atrativos e seguros.

Embora no início das pesquisas no tema, as cidades inteligentes tenham se limitado a uma abordagem puramente tecnológica (“*hard smartness*”), com ênfase nas tecnologias de informação e comunicação, muitas pesquisas atualmente mudam o foco para englobar, além da tecnologia bruta, os papéis do capital humano e educação, capital social e relacional e interesses ambientais (LOMBARDI, GIORDANO, *et al.*, 2012). Pardo e Nam (2011) definem uma cidade inteligente como uma combinação de três dimensões: tecnológica, humana e institucional, o que mostra que um grande valor deste conceito está na multidisciplinaridade.

Neste contexto, passou-se a discutir a implementação e uso de indicadores que permitissem avaliar o desenvolvimento do grau de inteligência das cidades levando em conta as diversas dimensões que as caracterizam. Entre esses modelos, SCMM (Modelo de maturidade de cidades inteligentes), WCCD (Conselho mundial de dados das cidades), o modelo brasileiro BR-SCMM e a ISO 37120, com indicadores para cidades sustentáveis (SANTANA e NUNES, 2021). Finalmente, em 2019, uma norma internacional foi publicada pela Organização Internacional para Padronização (ISO) visando especificar e estabelecer definições e metodologias para um grupo de indicadores de cidades inteligentes, a ISO 37122 (ISO, 2019). Nela estão presentes 85 indicadores divididos em 19 áreas temáticas, as quais se dividem em 6 domínios: economia inteligente, pessoas inteligente, governança inteligente, mobilidade inteligente, meio-ambiente inteligente e vida inteligente (SANTANA e NUNES, 2021). No ano seguinte ao lançamento da normativa internacional,

a Associação Brasileira de Normas Técnicas (ABNT) publicou uma versão traduzida para o português.

Entretanto, embora a publicação desta norma em português demonstre a vontade e a possibilidade de aplicá-la aqui, o contexto brasileiro ainda oferece alguns entraves. Um quarto da população brasileira ainda não tem acesso à internet e, destes, quase metade por inépcia (TOKARNIA, 2020). Ademais, embora o Brasil tenha uma Lei de Acesso à Informação e colete muitas informações macroscópicas em nível nacional pelos Censos decenais ou por pesquisas e acompanhamentos específicos, muitos dos dados presentes na ISO 37122 e necessários para o cálculo dos indicadores não são coletados ainda. Há de se notar, porém, que o país tem mecanismos para garanti-los e disponibilizá-los de forma segura e transparente, como prevê e exige a normativa.

Partindo de uma base de dados consolidada de indicadores e cidades, pode-se começar a inferir como outras cidades se irão comportar e, a partir disso, reduzir o trabalho de aquisição de dados. Por exemplo, em um caso onde todos os valores reais dos indicadores para uma determinada cidade também estão relacionados a um dado qualitativo (seja a classificação do valor do indicador em uma escala de “muito ruim” a “ótimo”), há algoritmos que usam de estatística e aprendizado de máquinas capazes de, dado o dado qualitativo, estimar o quantitativo. Também se poderia relacionar o grupo de cidades entre elas, criando subgrupos que permitam inferir informações relevantes ao conjunto.

A partir do agrupamento, se verificariam carências similares entre as cidades próximas, assim poder-se-ia criar um conjunto de intervenções necessárias e possíveis para este grupo, visando não o caso específico da cidade, mas o conjunto de cidades semelhantes criado, o que teria mais chances de alterar o cenário macro, além de ser de aplicação mais viável.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Aplicar ao contexto da avaliação de indicadores de cidades inteligentes reais, com base na ISO 37122 – Indicadores para Cidades Inteligentes e ISO 37120 – Indicadores para Serviços de Cidades e Qualidade de Vida, métodos de aprendizado estatísticos de definição de grupos e outros mecanismos úteis de forma a possibilitar um entendimento melhor dos grupos e das cidades agrupadas em níveis de inteligência.

Além disso, visa-se propor uma metodologia que permita, com base em análise estatística, reduzir o número necessário de indicadores a serem coletados para confecção dos grupos, provando ser capaz chegar aos mesmos agrupamentos utilizando todos os dados e apenas uma parte deles, barateando a análise.

2.2. OBJETIVOS ESPECÍFICOS

- Testar a validade da metodologia proposta com dados genéricos;
- Criar uma base de dados de cidades reais e seus indicadores com base nas ISO 37122 e/ou ISO 37120;
- Utilizar diferentes técnicas de aprendizado estatístico para gerar agrupamentos (PCA, dendrograma, etc);
- Com base na matriz de correlação entre indicadores (variáveis) e/ou nas construções dos *biplot* PCA, diminuir o número de indicadores de forma a manter os agrupamentos originais.

3. FUNDAMENTAÇÃO TEÓRICA E REVISÃO BIBLIOGRÁFICA

A sustentabilidade tem um papel muito grande na definição de cidades inteligentes. Evidência disto é que para referência normativa base do presente trabalho, em seu processo de aquisição de certificação, faz-se obrigatório o atingimento não só dos indicadores específicos para cidades inteligentes, na ISO 37122, mas também de sua norma-mãe, a ISO 37120 que versa especificamente sobre o tema de sustentabilidade (ISO, 2019).

Enquanto que no passado pensava-se a sustentabilidade puramente em se tratando do aspecto natural, de preservação de fauna e flora, eventualmente a palavra começou a ganhar sentido mais amplo. O desenvolvimento urbano sustentável passou a ser definido mais recentemente como sendo o balanço entre o desenvolvimento de áreas urbanas e a proteção do meio ambiente, com observância de equidade de renda, emprego, moradia, serviços básicos, infraestruturas social e de transportes nas áreas urbanizadas (AHVENNIEMI, HUOVILA, *et al.*, 2017).

O florescimento destas correntes de pensamento ambiental mais amplas veio junto do progresso dos grandes movimentos ambientalistas. O surgimento da agenda ambiental ocorreu na década de 70 com a primeira grande conferência voltada para o meio ambiente, a Conferência de Estocolmo de 1972. Os novos marcos ali definidos e a globalização cada vez mais acelerada com o fim da Guerra Fria acabaram intensificando a tendência de se quantificar todas as questões socioambientais (GUIMARÃES e FEICHAS, 2009).

O monitoramento urbano, entretanto, começou a ganhar vulto somente na última década do século passado em decorrência da Cúpula da Terra de 1992, no Rio de Janeiro, um evento que pavimentou o caminho das cidades para a sustentabilidade, propondo já indicadores para isso na forma da Agenda 21 (MARSAL-LLACUNA, COLOMER-LLINÀS e MELÉNDEZ-FRIGOLA, 2015). O Brasil aprovou sua Agenda 21 local em 2002. No mesmo ano, o IBGE publicou os Indicadores de Desenvolvimento Sustentável - IDS (MALHEIROS, PHILIPPI JR. e COUTINHO, 2008), contendo um conjunto de indicadores que auxiliariam a monitoração do progresso brasileiro no campo da sustentabilidade. A criação do conhecido Índice de Desenvolvimento Humano, IDH, aconteceu quase que concomitantemente à Conferência Rio-92.

Com o surgimento do conceito de cidades inteligentes ainda na década de 90, começou a se notar que o uso de tecnologias diversas auxilia comunidades a alcançarem suas metas de sustentabilidade. Este é o entendimento comum compartilhado pela Comissão Europeia, do continente que, desde 2012, apoia diversos projetos de cidade inteligentes

(AHVENNIEMI, HUOVILA, *et al.*, 2017). Com a maturação do conceito, cidades inteligentes começaram a agregar e a dar ênfase cada vez mais ao aspecto humano, a um ambiente que favoreça a aprendizagem, participação, pesquisa e desenvolvimento (CHIUSOLI e REZENDE, 2019).

Deixando de ser um conceito puramente técnico, cidades inteligentes já vêm promovendo, entre outras coisas, a geração de novas oportunidades de negócios. O Instituto Smart City Business America (SCBA), com sede em São Paulo, promete ser uma organização com visão orientada para negócios para geração de novas oportunidades neste ramo em torno do ecossistema de cidades inteligentes. Por meio de suas plataformas, o instituto oferece cursos de formação sobre cidades inteligentes, presta consultorias, promove fóruns, palestras e entrevistas. Por meio de iniciativas como estas, vê-se já o vulto que esta área começa a tomar e suas possibilidades de crescimento.

3.1. INDICADORES PARA CIDADES INTELIGENTES E SUSTENTÁVEIS

3.1.1. O conceito de indicador

Para começar a se aprofundar nas metodologias e conjuntos de indicadores, deve-se de antemão ter conhecimento do conceito-chave que serve de base para cada uma delas: o indicador, termo que vem do latim *indicare*, significando descobrir, anunciar, estimar. Em Magalhães (2004), encontra-se uma definição interessante para indicadores:

“[...] são, assim, parâmetros representativos, concisos e fáceis de interpretar que são usados para ilustrar as características principais de determinado objeto de análise”

Destaca-se assim a necessidade de os indicadores serem concisos, representativos e de fácil interpretação. Não só ilustrativos, no entanto, indicadores são utilizados como recursos de verificação de tendências ou fenômenos não detectáveis de forma imediata somente a partir dos dados brutos (SOBRAL, FREITAS, *et al.*, 2017).

Com base neste breve sumário, já se pode entender que nem todo dado colhido tornar-se-á automaticamente em um bom indicador. Ou seja, devem se ter critérios para a escolha dos indicadores que serão utilizados em algum determinado projeto ou monitoramento. Os mais relevantes critérios de seleção de indicadores ambientais sugeridos por Segnestam (1999) são: relevância direta para os objetivos do projeto, limitação em número, clareza no *design*, custos de desenvolvimento realistas, identificação clara de relações causais, alta

qualidade e confiabilidade, escala espacial e temporal apropriada, objetivos e linhas de base (que permitem a comparação entre momentos durante a interpretação dos resultados).

No mesmo documento publicado pelo Banco Mundial, Segnstam (1999) sugere que os indicadores de determinada metodologia estejam encaixados em quatro categorias:

- Indicadores de entrada (*input*): monitoram os recursos específicos de projeto fornecidos;
- Indicadores de saída (*output*): medem os bens e serviços fornecidos pelo projeto;
- Indicadores de resultado: medem os resultados imediatos ou de curto prazo da implementação do projeto;
- Indicadores de impacto: monitoram os resultados de longo prazo do projeto.

Embora alguns pesquisadores acabem distorcendo um pouco estas categorias, o cerne delas continua presente em todas as metodologias de classificação de indicadores, mesmo nas normas de padronização internacionais (ISO).

Pode-se afirmar que o objetivo da concepção de uma metodologia de indicadores é atingir, a partir destes, a informação e, por fim o conhecimento. Entende-se por informação como sendo o resultado das análises realizadas sobre um conjunto de indicadores, e por conhecimento a sistematização deste resultado.

Há também outra notação relevante que diferencia os indicadores de índices. Neste caso, índices seriam a agregação de dois ou mais indicadores; enquanto estes têm por objetivo indicar ou evidenciar um fenômeno, aqueles tentariam sinalizar por meio de um valor (chamado medida-síntese) a evolução de uma quantidade comparada a uma referência (SOBRAL, FREITAS, *et al.*, 2017).

Siche *et al.* (2007) desembaraçam ainda mais os conceitos. Usados como sinônimos em análises de mais baixo nível, índices e indicadores têm significados distintos: enquanto índices são valores numéricos que representam a realidade de um sistema simples ou complexo por meio de cálculos embasados cientificamente, indicadores são parâmetros selecionados e considerados isoladamente para refletir acerca apenas de condições do sistema analisado. Considera-se, assim, o índice como sendo um nível superior, surgindo da união de uma gama de indicadores ou variáveis. Para sair da conceituação abstrata, a expectativa de vida de uma população é um indicador, já o IDH (índice de desenvolvimento humano), como já informa o nome, é um índice.

Um empecilho muito claro e muitas vezes o mais relevante é o custo de obtenção dos indicadores. Caso os dados sejam de fácil coleta e obtenção (em portais de dados abertos ou

páginas da internet, por exemplo), o indicador se torna mais atrativo no quesito custo. Em casos de indicadores caros, com necessidade de implementação de tecnologia ou muito polimento dos dados brutos necessário, o indicador acaba se encarecendo e se tornando menos atrativo. Se ambos os casos, caro e barato, tenham a mesma relevância em relação ao fenômeno observado, é óbvio qual será escolhido.

O custo foi um dos grandes motivos alegados que levaram a maior coleta de dados gerais nacional, o Censo, a ser suspenso no ano de 2020 pelo governo federal, que não o incluiu no Orçamento (G1, 2021).

3.1.2. Conjuntos de indicadores e metodologias

Junto à consolidação dos conceitos de cidades inteligentes e cidades sustentáveis, um conjunto metodologias avaliativas foi proposto para que justamente fossem classificadas e acompanhadas cada uma das cidades em termos de inteligência, sustentabilidade, resiliência ou quaisquer outras lentes de análise. Diversos trabalhos foram publicados como tentativas para se medir esta suposta inteligência e sustentabilidade. Algumas metodologias acabam por focar mais em aspectos de infraestrutura bruta de telecomunicação, outros mais na avaliação do capital humano presente na cidade. Há também metodologias que se baseiam fortemente no sistema de transporte urbano (AHVENNIEMI, HUOVILA, *et al.*, 2017). O objetivo da criação de normas para realizar estas medições, especialmente normas transfronteiriças, é criar uma cultura de dados globalmente comparável e padronizada (SANTANA, OLIVEIRA NUNES e SANTOS, 2018).

Diversas metodologias, em especial as publicadas pelo Instituto Internacional de Padronização (ISO), contêm mais de um tipo de indicadores encontrados em seu corpo. Neste sentido, Huovila, Bosch e Airaksinen (2019), que realizaram um grande estudo comparativo entre as diversas metodologias internacionais de classificação e monitoramento de cidades inteligentes, descrevem uma taxonomia mostrada na Figura 3.1 para análise dos indicadores especificamente para este caso, estrutura esta que se mostrou verdadeiramente útil tanto na comparação entre as diversas normativas pesquisadas quanto na investigação de cada uma delas separadamente. Os pesquisadores inclusive propõem um tipo de indicador diferente dos quatro que foram anteriormente propostos pelo Banco Mundial, o indicador de processo. Este procura mensurar se as atividades planejadas realmente aconteceram.

Huovila, Bosch e Airaksinen (2019) além disso também propõem em qual estágio da implementação de medidas para criação de cidades inteligentes se deve implantar qual tipo de indicador sugerido em sua taxonomia.

Os grupos de indicadores analisados neste trabalho foram escolhidos com base em relevância, pioneirismo, inovação e relevância específica para o cenário brasileiro. Desta forma, mesmo conjuntos que não contemplavam nominalmente o conceito de cidades inteligentes, como os Indicadores de Desenvolvimento Sustentável do IBGE de 2002, acabam entrando no rol dos estudados devido não só à sua importância para o contexto nacional quanto também às interrelações entre os conceitos de cidades inteligentes e sustentabilidade, tema focal deste documento brasileiro, em todas as outras metodologias e normativas internacionais.

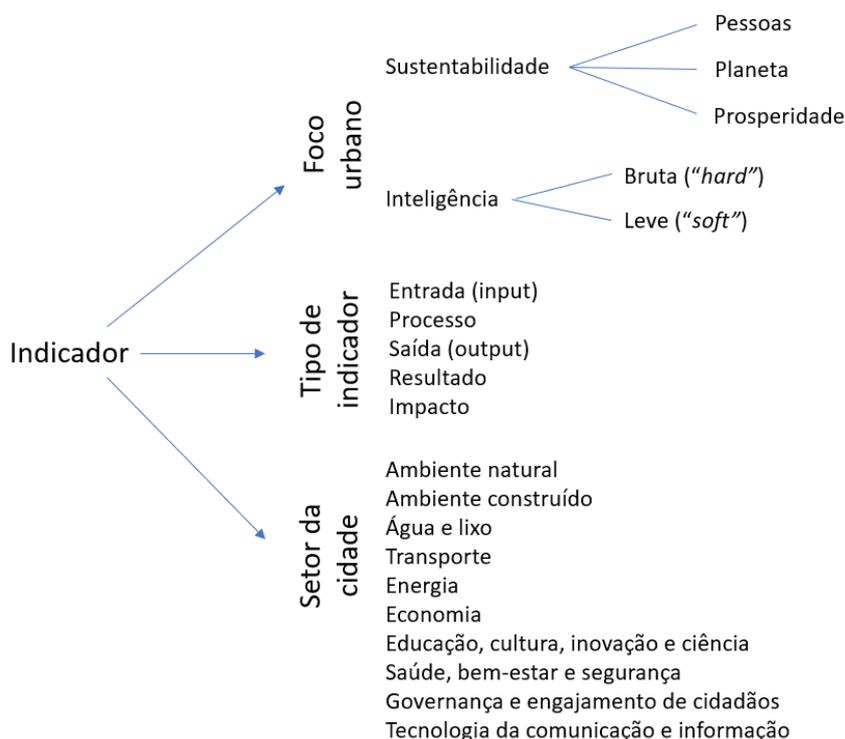


Figura 3.1 Taxonomia proposta por Huovila, Bosch e Airaksinen (2019)

Siche *et al.* (2007) admitem que não há uma fórmula ou receita que seja absolutamente consensual para avaliação do que é ou deixa de ser sustentável a partir dos indicadores. O mesmo se pode dizer acerca das cidades inteligentes, sendo este campo ainda mais de ponta que aquele.

3.1.2.1. Indicadores de Desenvolvimento Sustentável (2002)

Este conjunto de indicadores proposto pelo IBGE em congruência com a Agenda 21 Nacional aprovada no mesmo ano foi baseado na proposta de metodologia elaborada pela Comissão de Desenvolvimento Sustentável das Nações Unidas. Seu objetivo era

acompanhar a sustentabilidade do padrão de desenvolvimento do Brasil a partir de múltiplas dimensões (GUIMARÃES e FEICHAS, 2009).

Na primeira publicação dos indicadores, foram deixados de fora 15 dos 57 daqueles originalmente propostos pela ONU por não haver levantamento de dados regular, pela não existência de dados ou pela falta de relevância no contexto brasileiro. Entre eles, estavam a emissão de gases de efeito estufa, desertificação, população em assentamentos formais e informais, qualidade da água, entre outros. Na publicação seguinte, alguns indicadores faltantes foram incluídos, como os de desertificação, qualidade da água, adequação de moradia e acesso à internet. Há, entretanto, críticas acerca das lacunas deixadas na geração de informações ambientais por essa metodologia (MALHEIROS, PHILIPPI JR. e COUTINHO, 2008). Ao final das revisões, estavam presentes 60 indicadores divididos em 4 temáticas, como mostrado na Tabela 3.1 (KRAMA, 2009).

Por falta de união em torno deste conjunto de indicadores, esta referência acabou naufragando muito precocemente e não conseguiu surfar na onda ambientalista que varria o mundo com as conferências internacionais e o ganho de força do discurso verde no planeta.

Os indicadores propostos pelo IBGE ainda não incluíam nominalmente o conceito de cidades inteligentes e não as incluíam nem na revisão mais recente de 2008, mas foi esta estrutura de indicadores que inaugurou um modelo para acompanhamento urbano no Brasil em nível nacional. Também é importante ressaltar que não incluir nominalmente não significa deixar de abordar alguns dos temas concernentes à temática de cidades inteligentes. Na verdade, muitos dos indicadores presentes neste documento estarão ou presentes quase que integralmente em normas internacionais posteriores específicas da temática ou apenas levemente alterados, como os de percentual da população com acesso à internet.

3.1.2.2. European Smart Cities Ranking (2007)

Um dos primeiros sistemas robustos para avaliação de cidades inteligentes foi proposto por Giffinger et al. (2007). Ele tinha como objetivo, além de construir um ranking entre as cidades europeias, avaliar o desempenho delas, tornar-se uma possível ferramenta de auxílio a tomada de decisão, evidenciando pontos fortes e fracos de cada cidade, não meramente indicando as melhores e piores cidades (FLORES e TEIXEIRA, 2017). Visava-se, sobretudo, a criação de uma base de dados comum entre os diversos países e municipalidades europeus, o que facilitaria a troca de experiência entre os pares e criaria um potencial de aprendizado e melhoramento mais célere em diversas frentes.

Seis grandes eixos guias de indicadores guiaram este documento: economia inteligente, pessoas inteligentes, governança inteligente, mobilidade inteligente, meio ambiente inteligente, vivência inteligente. Estes eixos estariam presentes de alguma forma em diversas tentativas subsequentes de compor um corpo ideal de indicadores, inclusive na ISO de 2019.

Esta metodologia propôs a adoção de 64 indicadores, entre os quais PIB por pessoa empregada, habilidades em língua estrangeira, número de representantes municipais por residente, horas de sol e expectativa de vida. Vê-se, então, que se trata de uma gama ampla de indicadores, dos mais variados tipos, das mais variadas fontes, o que possibilita àquele que o usa para classificar uma cidade uma visão bem abrangente do local e de sua população.

Já em 2010, entretanto, o próprio autor evidenciou que a atenção da população ao European Smart Cities Ranking está concentrada principalmente nos rankings eles mesmos, totalmente negligenciando seu significado como instrumento de planejamento estratégico (GIFFINGER e GUDRUN, 2010). Então o intercâmbio de ideias esperado e aprendido conjunto acabou não acontecendo na escala esperada. Também, Flores e Teixeira (2017) argumentam que este sistema de indicadores acaba pecando na quantidade de indicadores ambientais, dando mais ênfase à convivência, economia e tecnologia. Este foi um dos grandes motivos pelos quais estes indicadores acabaram perdendo fôlego mais recentemente.

3.1.2.3. ISO 37120: Sustainable Development of Communities – Indicators for City Services and Quality of Life (2014)

Em 2014 foi publicada a primeira norma internacional da Organização Internacional de Normas (ISO) especialmente dedicada à sustentabilidade e qualidade de vida, a ISO 37120. Devido à reputação da entidade suíça, esta publicação acabou sendo extremamente relevante para o movimento de cidades inteligentes ao redor do mundo. Assim, esta normativa logo acabou sendo adotada por diversos países como diretriz local, a exemplo do Reino Unido e do Brasil, que publicaram a norma sob os selos de BS (padrão britânico) e NBR (norma brasileira). Seguindo as melhores práticas, a ISO também tem a cultura de manter suas normas sempre atualizadas, evidenciado no fato de que uma nova revisão do documento foi lançada em julho de 2018 contendo algumas alterações, remoções e inclusões de indicadores (HUOVILA, BOSCH e AIRAKSINEN, 2019).

Reconhecendo a necessidade de indicadores para o acompanhamento de performance e progresso, este documento objetiva ajudar as cidades na medição e manutenção de serviços da cidade e qualidade de vida ao longo do tempo e permitir compartilhamento e comparação de medidas entre cidades, gerando aprendizado mútuo (ISO, 2014). Este documento também

está organizado ao redor de eixos temáticos para basear seus indicadores, mas mais especificamente que aqueles presentes na European Smart Cities Ranking de 2007. No total, foram apresentados 104 indicadores distribuídos em 19 eixos que estarão novamente presente na Norma de cidades inteligentes. Estas categorias estão mostradas na Tabela 3.1. Há três tipos de indicadores presentes: indicadores essenciais, que devem ser seguidos na implementação da Norma impreterivelmente, indicadores de apoio, que são convenientes de serem seguidos e indicadores de perfil, cujo objetivo é caracterizar a cidade (RIBEIRO, 2019).

Além de definir os indicadores, a norma também contém as indicações como calculá-los, como considerar casos parciais, fontes de dados, todas as definições necessárias além das definições de como interpretar os dados.

Após a publicação deste documento, criou-se o Conselho Mundial de Dados de Cidade (WCCD) visando facilitar a aplicação da ISO pelo mundo e criando uma plataforma de comparação entre as cidades aderidas (RIBEIRO, 2019). Além disso, o WCCD também certifica as cidades e garante a elas uma de 5 classificações, baseadas no número de indicadores seguidos. Um fato importante é que esta norma não traz classificações no tocante ao valor dos indicadores, nem dá um valor mínimo a ser atingido, tornando o trabalho de WCCD importante no sentido de indicar com precisão quais cidades são de fato exemplares e devem ser tomadas como autoridade nas frentes tratadas pela normativa. O banco de dados do WCCD não esteve acessível durante toda a duração da pesquisa que resultou nesta monografia.

No Brasil, a NBR ISO 37120 surgiu em 2017 da falta de uma ferramenta consistente para avaliar a sustentabilidade urbana. Algumas cidades já realizam a coleta de alguns indicadores desta Norma visando a obtenção de certificação ou não, Brasília é uma delas.

3.1.2.4. ETSI TS 103 463: Key Performance Indicators for Sustainable Digital Multiservice Cities (2017)

Este documento publicado pelo Instituto de Padronização em Telecomunicação Europeu (ETSI) em 2017 se propõe a ser relevante para o mesmo escopo de cidades que o European Smart Cities Ranking de 2007 abordado anteriormente. Nota-se, entretanto, já uma relevância maior de aspectos de inteligência bruta. Na introdução do documento, lembra-se do que usuários de tecnologias de comunicação e informação esperam atingir: baixas emissões de carbono, cidades eficientes em seu uso de recursos, facilitação de comunicação

entre pares de uma cidade, criar confiança em soluções propostas, monitorar o progresso, tudo por meio de uma estrutura metodológica comum (ETSI, 2017).

Este documento contém 76 indicadores divididos em quatro grandes categorias: pessoas, planeta, prosperidade e governança/gestão. Entre as referências utilizadas para composição desta normativa está a ISO 31720 (2014), já apresentada neste trabalho. Uma novidade deste conjunto de indicadores está na presença de alguns diretamente relacionados à segurança cibernética, algo ausente inclusive na norma ISO mais recente de indicadores para cidades inteligentes. Imagina-se que, devido aos recentes escândalos envolvendo grandes monopólios de tecnologia, em sua vasta maioria estadunidenses, e a privacidade dos usuários, além dos também recentes embates entre governos e estas corporações, indicadores acerca de segurança cibernética e privacidade na *web* serão de crucial importância em normativas relevantes futuras.

3.1.2.5. ISO 37122: Sustainable Cities and Communities – Indicators for Smart Cities (2019)

Já em 2014, a ISO visava criar uma nova série de padrões normativos para estabelecer uma abordagem integrada e holística para o desenvolvimento sustentável e resiliência quando publicou a ISO 37120 (ISO, 2014). Visando dar mais um passo na entrega deste objetivo, publicou a ISO 37122 voltada especificamente para cidades inteligentes. Logo no início do documento, uma cidade inteligente é definida como uma cidade que aumenta o ritmo no qual ela provê resultados sociais, econômicos e de sustentabilidade ambiental, além de responder a desafios como a mudança climática, célere crescimento populacional e instabilidade política e econômica (ISO, 2019). Ou seja, já na definição do objeto tema, traz-se à tona o conceito de ritmo, evidenciando a utilidade dos indicadores sugeridos.

Este documento é composto de 85 indicadores agrupados em 19 eixos temáticos. Uma das grandes diferenças em relação à norma anterior de 2014 é que esta adotou o conceito de Internet das Coisas (IoT) na aquisição de dados das áreas de resíduos sólidos, transporte e consumo de água, entre outras que este tipo de aquisição era viável (HUOVILA, BOSCH e AIRAKSINEN, 2019). Esta mudança, aparentemente simples, acaba atualizando e dinamizando a aquisição de indicadores, além de encarecê-los.

Uma análise interessante trazida por Houvila, Borsch e Airaksinen (2019) trata do balanço entre indicadores que evidenciam sustentabilidade e aqueles que evidenciam inteligência, totalmente condizente com a taxonomia proposta por eles mostrada na Figura 3.1. Concluiu-se que a Norma ISO de 2014 era composta de 89% de indicadores de

sustentabilidade, enquanto a mais recente, de 2019, de apenas 36%; ou seja, 64% de seus indicadores eram de inteligência. Também se notou que o documento de 2014 focava mais em sustentabilidade das pessoas e em inteligência leve (“*soft*”), atributos não tangíveis da inteligência, enquanto a Norma mais recente focava mais em sustentabilidade do planeta e inteligência bruta, como atributos tangíveis, infraestrutura, tecnologia da informação e comunicação, entre outros.

Também é interessante notar a preocupação dos formuladores desta normativa às metas já estabelecidas de sustentabilidade da Organização de Nações Unidas em 2015. Neste ano, a ONU publicou um documento que definia 10 objetivos a serem almejados pelos países membros, dentre os quais: a erradicação da pobreza, erradicação da fome, garantia de igualdade em todas as suas formas, garantia de energia sustentável para todos, construção de infraestrutura resiliente e redução das desigualdades entre os países. As metas arrojadas impostas pelo organismo multilateral estão presentes no Anexo B da Norma de 2019 (ISO, 2019). Lá, estão relacionados alguns dos indicadores presentes no documento a cada uma das metas de 2015. Um exemplo é a relação entre o décimo objetivo, de reduzir desigualdades dentro e entre países, aos indicadores 13.2, que trata da percentagem do orçamento municipal alocado para provisão de ajudas de mobilidade para cidadãos portadores de necessidades especiais (PNE) e 13.4, que trata da percentagem do orçamento municipal alocado para provisão de programas que visam acabar com a segregação digital. Esta adequação evidencia que este conjunto de indicadores é muito mais integrado que os outros aqui já apresentados e tem um escopo muito mais amplo e sólido de aplicação.

Por se tratar de uma Norma muito recente e com indicadores muito distintos dos que já estão consolidados na área no Brasil, apenas uma cidade brasileira que se tem notícia tem tentado obter a certificação, São José dos Campos em São Paulo. Uma crítica comum a esta normativa é justamente a dificuldade de obtenção dos dados requeridos.

A Tabela 3.1 a seguir apresenta um apanhado básico dos documentos tratados até aqui.

Tabela 3.1 Comparação entre normativos de indicadores para cidades inteligentes e sustentáveis em suas revisões mais recentes

Nome do documento	Categorias principais	Número de indicadores
Indicadores de Desenvolvimento Sustentável (revisão 2008)	Ambiental, social, econômica e institucional. ¹	60 ³

Nome do documento	Categorias principais	Número de indicadores
European Smart Cities Ranking (2007)	Economia inteligente, pessoas inteligentes, governança inteligente, mobilidade inteligente, meio ambiente inteligente, vivência inteligente. ²	64 ²
ISO 37120 Sustainable Development of Communities – Indicators for City Services and Quality of Life (revisão 2018)	Economia, educação, energia, ambiente e mudança climática, finanças, governança, saúde, habitação, população e condições sociais, recreação, segurança, resíduos sólidos, esporte e cultura, telecomunicação, transportes, agricultura local e urbana e segurança alimentar, planejamento urbano, água e esgoto. ³	104 ¹
ETSI TS 103 463: Key Performance Indicators for Sustainable Digital Multiservice Cities (2017)	Pessoas, planetas, prosperidade e governança. ³	76 ¹
ISO 37122 Sustainable Cities and Communities – Indicators for Smart Cities (2019)	Economia, educação, energia, ambiente e mudança climática, finanças, governança, saúde, habitação, população e condições sociais, recreação, segurança, resíduos sólidos, esporte e cultura, telecomunicação, transportes, agricultura local e urbana e segurança alimentar, planejamento urbano, água e esgoto. ³	85 ¹

¹Krama, 2009 ; ²Giffinger *et al.*, 2007; ³Huovila, Bosch e Airaksinen, 2019.

3.2. AGRUPAMENTOS ESTATÍSTICOS

Para facilitar as análises de um grupo heterogêneo de dados, diversas metodologias estatísticas podem ser empregadas. A formação de grupos hierarquizados é uma destas e é extremamente útil (NIELSEN, 2016). Há, entretanto, mais de uma forma de agrupar os dados a partir de uma base estabelecida e também diversas formas de apresentar as conclusões.

Ressalta-se que não só agrupamentos hierárquicos são úteis. Há diversas técnicas que envolvem ou não o aprendizado estatístico que são capazes de gerar agrupamentos e reduzir drasticamente a dimensionalidade dos dados iniciais sem perda significativa de informação que geram agrupamentos de enorme valia, mas não hierarquizam os dados. Algumas destas são apresentadas neste trabalho.

3.2.1. Dendrograma

A palavra dendrograma vem do grego *dendron*, que significa “árvore” e *gramma*, que significa “desenhar”. Trata-se de uma representação de divisões binárias organizadas em forma de árvore (NIELSEN, 2016, p. 223). O dendrograma consiste em várias linhas em formato de U que conectam pontos de dados em uma árvore hierárquica, com a altura de

cada U representando a distância entre os dois pontos de dados conectados. Esta técnica diferente de outras que serão aqui apresentadas é uma divisão hierárquica, o que quer dizer que ela não somente mostra os graus de similaridade entre os pontos dados, mas também define claramente as conexões entre eles, hierarquizando o resultado.

Um exemplo de dendrograma está mostrado na Figura 3.2, onde se vê uma grande similaridade entre as observações 2 e 10, uma similaridade parecida entre as duplas 5, 8 e 6, 7; mas uma distância muito grande entre 6 e 2, por exemplo. Nesta figura, os grupos definidos poderiam ser (2,10), (5,8,9), (1,4) e (3,6,7). A tolerância do distanciamento entre os pontos de cada grupo, entretanto, pode acabar gerando disposições diferentes.

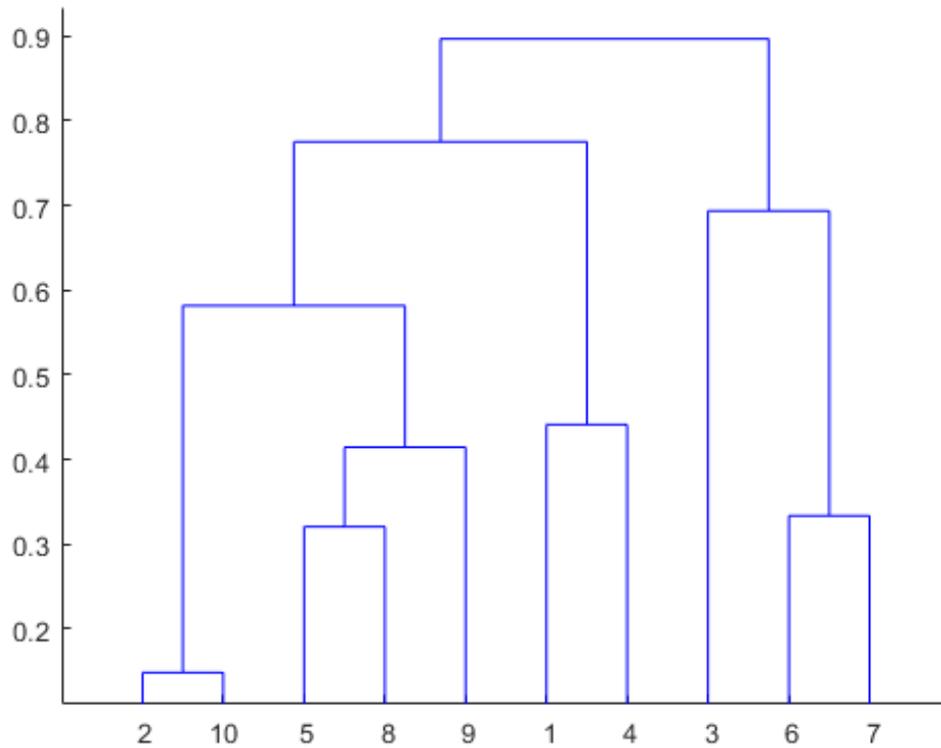


Figura 3.2 Exemplo de dendrograma (MATLAB)

Várias métricas de distância podem ser usadas para a confecção de um gráfico como este. A mais simples e que será utilizada neste trabalho para esta técnica será a distância euclidiana, dada pela Equação 3.1:

$$d_{ab} = \|a - b\| = \sqrt{\sum_i (a_i - b_i)^2} \quad \text{Equação 3.1}$$

sendo a e b as observações e a_i e b_i suas características. Um exemplo clássico entre as bases de dados já conhecidas é o de modelos de carros (a) e suas características de motor, potência, número de marchas, entre outros (b).

Em um gráfico n -dimensional com as características compondo seus eixos, a distância euclidiana é exatamente a distância entre os pontos neste espaço, sem nenhuma transformação.

Um algoritmo genérico de funcionamento de um código gerador de um dendrograma, como descrito por Hintze (2007, p. 187-188), é:

Seja a distância entre os grupos i e j ser representada por d_{ij} com o grupo i contendo n_i objetos. Seja D o conjunto de todos os d_{ij} restantes. Supõe-se que há N objetos a agrupar.

1. Encontrar o menor elemento d_{ij} restante em D .
2. Juntar os grupos i e j em um novo grupo, k .
3. Calcular um novo conjunto de distâncias d_{km} usando a fórmula

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma |d_{im} - d_{jm}| \quad \text{Equação 3.2}$$

Com m representando qualquer outro grupo que não k . Essas novas distâncias substituem d_{im} e d_{jm} em D . Também seja $n_k = n_i + n_j$. Os coeficientes α_i , α_j , β e γ são escolhidos de acordo com o algoritmo específico utilizado. No caso do presente trabalho, usou-se $\alpha_i = \alpha_j = 0,5$, $\beta = 0$ e $\gamma = -0,5$.

4. Repetir os passos 1 a 3 até que D contenha um só grupo ao qual pertençam todos os objetos. Isso requererá $N-1$ interações.

Além disso, dendrogramas podem ser apresentados na horizontal, vertical, com ou sem cores entre os ramos da árvore. Ademais, em alguns casos, pode-se provar útil não utilizar a matriz real dos dados, mas alguma matriz normalizada, o que poderia corrigir distorções e facilitar a interpretação dos resultados.

3.2.2. Escalonamento Multidimensional (MDS)

O escalonamento multidimensional é uma representação visual de distâncias ou dissimilaridades entre um conjunto de objetos (KRUSKAL, 1978). Estes podem se referir a

medições de folhas de diversas espécies, cidades, coordenadas, orientação política ou um conjunto de indicadores.

Trata-se de uma técnica capaz de reduzir a dimensionalidade do problema tratado criando um gráfico de posições relativas de um número de objetos dada uma matriz de distâncias euclidianas entre eles, sendo inclusive a redução de dimensionalidade atingida o maior valor da técnica. O termo “multidimensional” faz referência ao fato de que o resultado gerado pode ter qualquer valor de uma a n dimensões. Normalmente opta-se por gráficos de duas ou três dimensões, já que valores maiores começam a dificultar a visualização dos resultados. Nota-se que a Equação 3.1, como uma extensão do Teorema de Pitágoras, é capaz de calcular a distância independentemente do número de dimensões de um determinado conjunto de dados.

As duas técnicas mais conhecidas de escalonamento multidimensional são o escalonamento multidimensional métrico (mMDS), e a técnica não-métrica (nMDS). A diferença primordial entre eles é que o primeiro algoritmo assume que a transformação que ocorreu na matriz inicial de dados é linear. Uma destas transformações possíveis é a preconizada pela Equação 3.1 (WANG, 2012, p. 117-118). Ou seja, a matriz distância é dada por:

$$D = [D_{ij}] = [d(x_i, x_j)]_{i,j=1}^n \quad \text{Equação 3.3}$$

Assim, pode-se inferir que a função que define o método mMDS é explicitamente definida como linear, enquanto que no caso do nMDS, as distâncias se apresentam em escala ordinal, ou seja, relativa (de primeiro a último); sendo também a única fonte confiável de proximidades. Assim, o que antes era uma relação linear, passa a ser definido por uma série de curvas, às vezes simples, mas por vezes complexas (GLEN, 2015). Também é possível se utilizar uma mistura de ambas as técnicas em casos específicos.

Para o caso de dados de indicadores contáveis e não-ordinais de cidades, é preferível adotar-se o método métrico, mMDS.

O algoritmo de funcionamento leva em conta autovalores e autovetores para realizar o escalonamento. Autovetores são os vetores que não mudam de direção após sofrerem uma transformação linear, mas apenas são modificados por algum escalar, chamado autovalor. Percebe-se, então, que a ideia por trás de os usar em escalonamentos multidimensionais é atingi-lo sem gerar distorções.

O uso desta técnica é recomendado por Kruskal (1978), autor de um dos livros seminais acerca do escalonamento multidimensional, em situações que são dadas um conjunto de distâncias e diferenças e o objetivo é criar um mapa de onde cada ponto está localizado. Cita-se o exemplo clássico de que, caso se tenha a posição exata de um grupo de cidades, a técnica se torna inútil, mas no caso de só se ter como dados a matriz de distâncias entre as cidades, o mMDS é capaz de gerar uma distribuição espacial delas, facilitando a interpretação dos dados brutos.

O algoritmo genérico, como exposto em Hintze (2007, p. 145), é dado por:

Seja a matriz de distâncias euclidianas D calculada de uma configuração de objetos X como na Equação 3.3.

1. A partir de D , calcular A , com

$$A = \left\{ -\frac{1}{2} d_{ij}^2 \right\} \quad \text{Equação 3.4}$$

2. A partir de A calcular B , com

$$B = \{ a_{ij} - a_{i.} - a_{.j} + a_{..} \} \quad \text{Equação 3.5}$$

onde $a_{i.}$ é a média de todos os a_{ij} através de j .

3. Encontrar os p maiores autovalores $\lambda_1 > \lambda_2 > \dots > \lambda_p$ de B e autovetores correspondentes $L = (L_{(1)}, L_{(2)}, \dots, L_{(p)})$ que são normalizados tal que $L'_{(i)} L_{(i)} = \lambda_i$. Assume-se que p é escolhido de tal forma que os autovalores sejam todos relativamente grande e positivos.
4. As coordenadas dos objetos são as linhas de L .

De forma a compensar potenciais distorções, ao invés de usar os valores reais, medidos, como dados de D , pode-se normalizá-los a partir de alguma transformação linear como também dito para o dendrograma.

A similaridade desta técnica e dos dendrogramas são que ambas permitem a visualização de relações, mas não apresentam os dados da mesma maneira. Diferentemente dos dendrogramas, MDS não define os agrupamentos em estruturas hierárquicas.

A leitura do gráfico gerado se dá de forma extremamente intuitiva: caso os pontos estejam próximos uns dos outros formando uma nuvem, trata-se de um grupo similar; caso

estejam afastados, trata-se grupos diferentes. Quanto mais próximo, mais similar. A Figura 3.3 a seguir exemplifica um destes gráficos em uma aplicação muito comum.

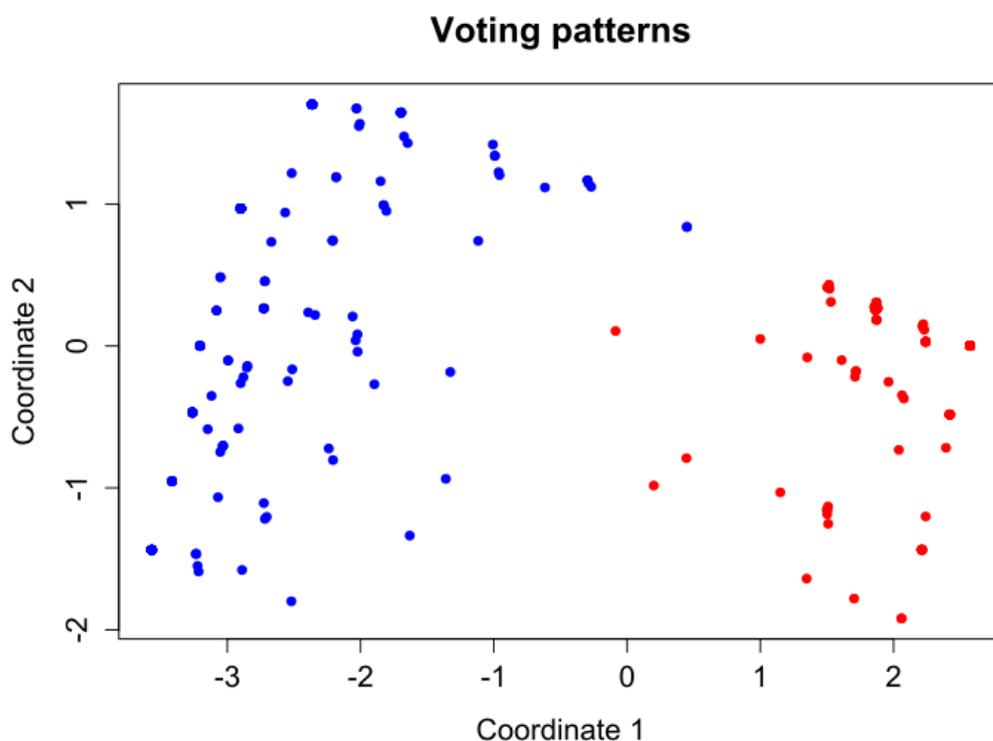


Figura 3.3 Escalonamento Multidimensional clássico aplicado sobre dados de votos de membros do Congresso dos Estados Unidos (ANDREWMAN327, 2013)

No caso da Figura 3.3, foi atribuído um valor de 1 para votos “Sim”, -1 para votos “Não”, e votos de abstenção receberam o valor 0, em um espaço amostral de 10 projetos de lei a passar o congresso. Os pontos azuis são do partido Democrata e os pontos vermelhos, do partido Republicano. Pode-se notar o poder e a valia da técnica na separação de grupos distintos e espacialização de dados, além de evidenciar o poder de transformação de um conjunto bruto de dados (neste caso uma matriz de número de deputados pelo número de leis escolhidas como amostra) em um gráfico de fácil interpretação.

3.2.3. Análise de Componente Principal (PCA)

A análise de componente principal é uma técnica multivariada que analisa tabelas de dados que são descritos por diversas variáveis quantitativas dependentes e interrelacionadas, realizando uma análise de similaridade entre as amostras. Diferentemente da técnica anterior, esta consegue, além de criar uma distinção de grupos de pontos, manter uma visualização das interações de parâmetros. Esta metodologia objetiva extrair as informações mais importantes dos dados, comprimir o tamanho do conjunto de dados, mantendo somente as

informações mais importantes, simplificar a representação e analisar a estrutura das observações e variáveis. Para tal, representam-se os padrões de similaridade entre as observações e as novas variáveis ortogonais capazes de indicar relações entre os parâmetros (ABDI e WILLIAMS, 2010).

É uma técnica poderosa de achatamento dimensional de dados, uma vez que permite manter consistentemente as relações entre as variáveis intactas. Se os dados são bem aproximados pelas suas duas componentes principais, o *biplot* gerado em uma PCA permite a visualização de dados de altas dimensões em apenas duas.

Ringnér (2008) demonstra o uso da PCA para análise de dados em estudos genéticos nos quais a abundância de dimensões a serem analisadas criam um impasse à visualização padrão dos dados. Nestes casos, como no caso analisado no presente trabalho, a redução dimensional torna-se imperativo para boas análises. No caso estudado por ele, de uma base de dados original com 8534 variáveis de 105 amostras, as duas componentes principais retiveram 22% da variância original; 63 componentes eram necessárias para reter 90% dela e a variância original era finalmente atingida com 104 componentes principais, uma redução enorme das mais de 8000 variáveis iniciais. O autor conclui que, quando o número de variáveis é maior que o número de amostras, a PCA pode reduzir a dimensionalidade das amostras para, no máximo, o número de amostras sem incorrer em perda de informação.

Torres-Salinas, Robinson-García *et al.* (2013) utilizam esta técnica para uma análise bem similar à realizada no presente trabalho. Eles comparam os 21 países europeus (à época) com base em 8 indicadores de pesquisa e desenvolvimento. Com apenas duas componentes principais, atingiu-se um valor de 99,9% da variância original, ou seja, as amostras que eram antes de 8 dimensões passaram a ser exibidas em um gráfico de apenas 2 com perda quase nula de informação. O mesmo procedimento foi aplicado pelos autores para uma comparação entre as 25 melhores universidades do planeta (de acordo com o ranking THE), e também para os 12 campos científicos da Universidade de Granada, Espanha, de acordo com indicadores bibliométricos. A cada exemplo posterior, a perda de variância com apenas duas dimensões foi ficando mais significativa. Os autores concluem, enfim, reconhecendo que esta ferramenta é útil não só pelas análises dos dados multivariados, mas também pela sua apresentação fácil de ser entendida.

Como no método anterior de MDS, os cálculos desta técnica também se baseiam em autovalores e autovetores para fazer a regularização e redução dimensional dos dados.

Um dos algoritmos genéricos simplificados usados para realizar a análise de componente principal, baseado na covariância, está mostrado a seguir.

Seja a matriz X de dados com linhas representando itens e colunas representando características.

1. Criar Z , a matriz normalizada de X a partir da equação

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad \text{Equação 3.6}$$

com μ sendo a média dos valores da coluna j e σ_j o desvio padrão da coluna j .

2. Calcular a matriz covariância de Z a partir de sua transposta, como mostrado em

$$COV(Z) = Z^T Z \quad \text{Equação 3.7}$$

3. Calcular a matriz de autovetores P e matriz diagonal D de autovalores a partir da decomposição

$$Z^T Z = P D P^{-1} \quad \text{Equação 3.8}$$

4. Organizar a matriz P em ordem de maior para o menor autovetor e chamar o resultado de P^* .
5. Calcular novas características Z^* , matriz onde cada observação é uma combinação das variáveis originais e os pesos definidos pelos autovetores. Como os autovetores são independentes, também serão as colunas de Z^* .

$$Z^* = Z P^* \quad \text{Equação 3.9}$$

6. Finalmente, selecionar o número de dimensões desejadas.

Espera-se obter com isso uma distribuição dos pontos estudados e alguns vetores indicando a relação ou não entre os parâmetros estudados (colunas da matriz X). Por um mesmo gráfico mostrar ambas as coisas, ele é conhecido como *biplot*, com “bi” indicando a duplicidade no mesmo. Os grupos formados indicam proximidade entre seus parâmetros com base na proximidade entre eles, semelhante ao que ocorre no escalonamento multidimensional. Os vetores criados, entretanto, indicam o seguinte: quando dois deles estão muito próximos um do outro (ou seja, o ângulo entre eles é pequeno), indica forte relação positiva entre os dois parâmetros; quando estão ortogonais ou próximos disso, indica a ausência de relação entre os parâmetros; e, caso estejam diametralmente opostos, são negativamente relacionados. Além disso, o tamanho de cada um dos vetores indica o desvio padrão da característica que ele evidencia para as componentes principais escolhidas.

Para se verificar a validade da visualização com menos dimensões, pode-se analisar a perda de variância (aqui representando a informação das amostras originais) para o número de componentes principais analisadas. Um gráfico possível de se construir é o *scree plot*, que indica exatamente o quanto da variância original é mantida para cada uma das PC (componentes principais) mantidas. Um exemplo de um destes gráficos está mostrado na Figura 3.4.

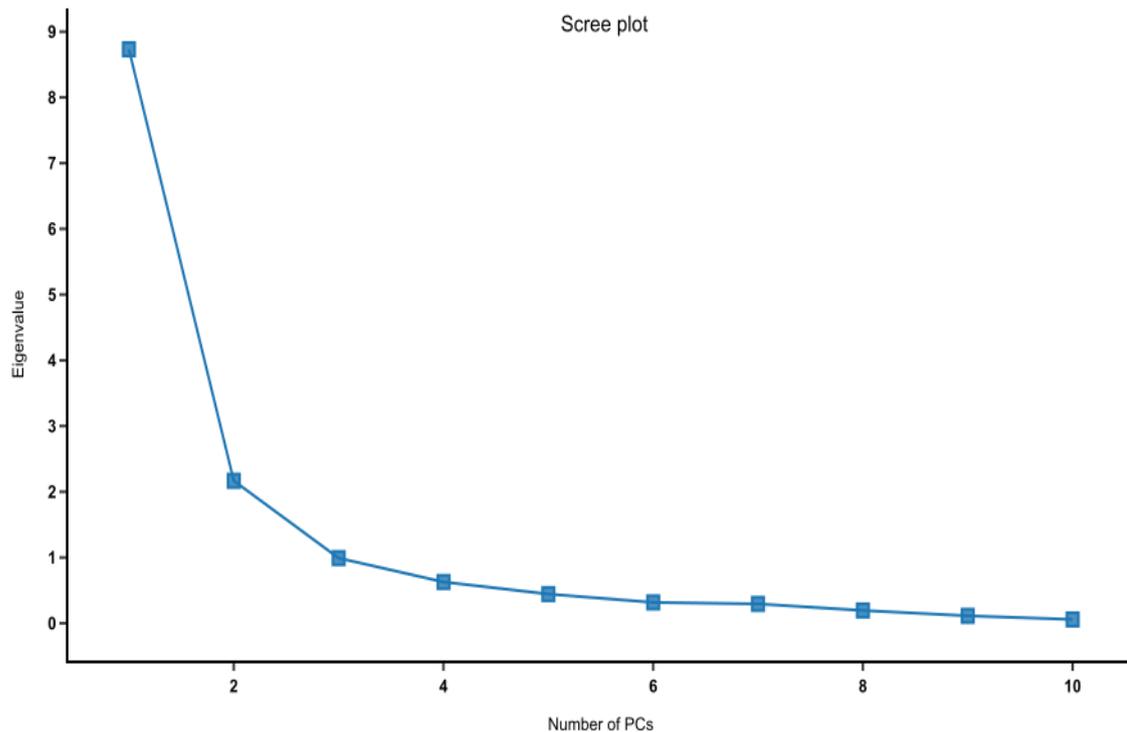


Figura 3.4 Exemplo de *Scree plot* de uma Análise de Componente Principal (NGO, 2018)

Na Figura 3.4, embora o eixo y diga autovalores, neste caso, eles essencialmente indicam a quantidade de variação nos dados. Usa-se desses gráficos para se escolher o número de PCs a se utilizar. Idealmente, um *scree plot* teria uma curva em “L”, ou seja, com uma curva abrupta após dois ou três valores, e uma estabilização próxima de zero. Isso indicaria que apenas estes três primeiros valores gerariam uma visualização ótima dos dados originais. No caso da Figura 3.4, duas ou três componentes principais seriam já boas para gerar uma visualização fidedigna aos dados originais no *biplot* PCA.

Supondo que apenas duas dimensões fossem julgadas como necessárias para a composição do *biplot*, ou seja, o gráfico gerado seria um gráfico plano, composto das nuvens de pontos e vetores, esta visualização seria algo conforme o mostra na Figura 3.5.

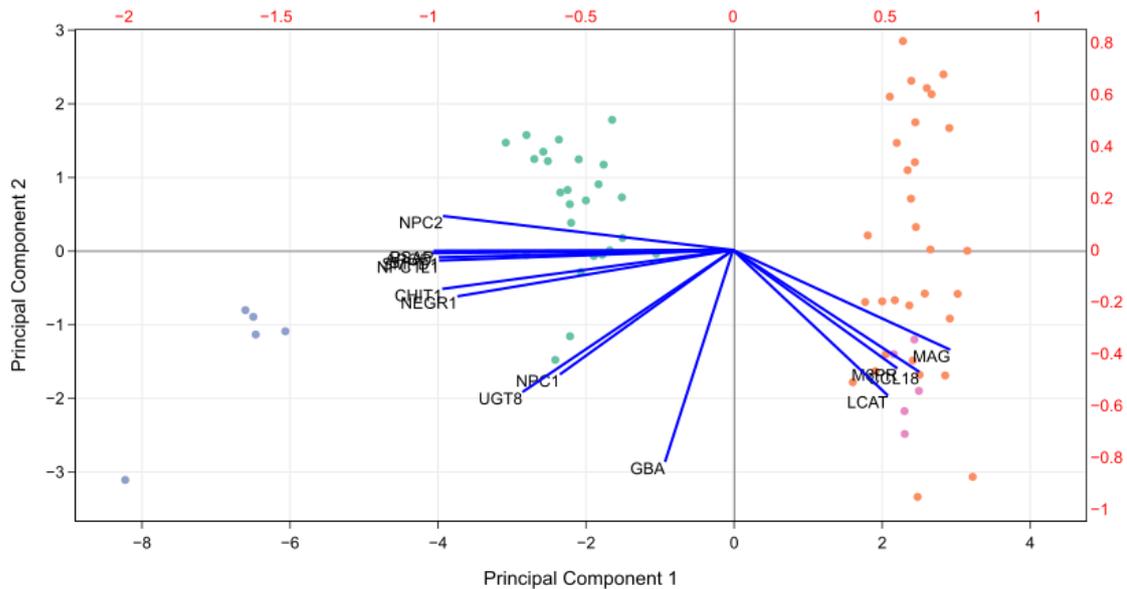


Figura 3.5 Exemplo de *biplot* PCA (NGO, 2018)

No caso da Figura 3.5, vê-se que MAG e NPC2, dois parâmetros dos dados analisados, estão negativamente relacionados. Pode-se dizer que a relação entre MAG e GBA não é muito evidente e que a relação entre NPC1 e UGT8 é fortemente positiva. Também se percebe a formação de três grupos bem distintos de pontos.

Vale ressaltar que uma diferença primordial entre esta técnica e o MDS é que aquela usa uma matriz de similaridade para gerar os resultados, enquanto que esta usa os dados originais. Também se ressalta que a PCA, assim como o MDS, não realiza um agrupamento hierárquico das amostras como no caso do dendrograma.

4. METODOLOGIA E ANÁLISES INICIAIS

4.1. APRESENTAÇÃO E DEFINIÇÕES

A metodologia consiste em aplicar os diversos métodos de agrupamento hierárquico definidos em 3.2 e interpretar os resultados obtidos. Para melhor ilustração, optou-se por fazer um exemplo com dados fictícios e realizar uma aplicação em pequena escala das técnicas propostas, assim evidenciando o poder do método. Para tal, foram criadas 12 cidades (A, B, C, ..., L) e 10 indicadores para cada uma delas. Nesta fase, a ISO 37122 foi apenas soltamente sendo usada como base, sendo escolhidos números de indicadores fictícios condizentes com as subcategorias da Norma, três de economia, três de gestão ou governança, dois de energia, e dois de transportes. Os indicadores utilizados com base e seus respectivos títulos autoexplicativos estão mostrados na Tabela 4.2. Todos os valores gerados foram algum valor de percentagem, de 0 a 100. Os resultados simulados estão expostos na Tabela 4.1.

Tabela 4.1 Conjunto de dados de cidades fictícias e seus indicadores simulados

	Ind. 5.1	Ind. 5.2	Ind. 5.3	Ind. 10.1	Ind. 10.2	Ind. 21.1	Ind. 7.6	Ind. 7.9	Ind. 19.4	Ind. 19.5
A	82.4	63	67.2	68.3	75.1	77.1	71.1	78.6	71.1	84.7
B	71.5	72.8	73.6	78.6	84.3	86.9	87	63.8	67.6	79.5
C	55.4	56.8	61	10	62.3	3.5	12.4	12.8	66.8	59.1
D	71.2	65.1	21.8	38.5	41.1	25.9	11.9	12	84.8	15.8
E	82.9	72.9	79.3	63.2	86.5	73.9	84.4	82.1	68.8	71
F	70.3	75	46	10.7	10.4	64	8.8	9.4	52.8	93.9
G	22.3	34.9	14.4	74.5	88.7	40.8	0.7	13.5	17.5	16.2
H	22	6.2	23.5	8.7	5.6	13.6	12.7	5	6.4	7.7
I	29.1	2.6	3	18.1	8	18.3	19.7	5.9	2.3	11.9
J	26	23.2	8.8	28.9	23.9	1.8	21.7	9.5	27.4	2.2
K	26	23.2	8.8	28.9	23.9	18.3	19.7	5.9	2.3	11.9
L	13	12.2	4.8	14.9	12.9	0.8	10.7	4.5	13.4	1.2

Os títulos das colunas da Tabela 4.1 mostram quais os indicadores propostos pela ISO 37122 foram escolhidos como exemplo o atual exercício. Ressalta-se que na Norma há diversos indicadores propostos que não se enquadrariam nos dados de percentagem aqui apresentados; este, entretanto, não seria um entrave à aplicação da metodologia, já que a formação dos grupos independe desta condição.

Tabela 4.2 Indicadores ilustrativos selecionados

	Número do Indicador	Conteúdo
Economia	5.1	Percentagem dos contratos de serviços de cidades que contém política de dados abertos
	5.2	Taxa de sobrevivência de novos negócios por 100.000 habitantes
	5.3	Percentagem da força de trabalho em ocupações na área de tecnologia de comunicação e informação
Gestão	10.1	Número anual de visitas online ao portal municipal de dados abertos
	10.2	Percentagem dos serviços da cidade acessíveis e que podem ser pedidos online
	21.1	Número anual de cidadãos partícipes no processo de planejamento por 100.000 habitantes
Energia	7.6	Percentagem da iluminação pública controlada por sistema de performance
	7.9	Percentagem dos edifícios com medidores de energia inteligentes
Transporte	19.4	Número de bicicletas disponíveis por serviço disponibilizado pela municipalidade por 100.000 habitantes
	19.5	Porcentagem de linhas de transporte público equipada com sistema de tempo real acessível publicamente

A Tabela 4.1 foi criada com uma mistura de dados aleatórios e de dados manuais. Para se obter clareza na distinção dos grupos, os indicadores das duas primeiras cidades (A e B) foram criados obedecendo uma distribuição uniforme com limite inferior de 60 e superior de 90, garantindo que todos os valores seriam relativamente altos, mas não irrealistas demais. As próximas duas cidades, C e D, foram criadas com um intervalo mais amplo, mas de média inferior; o que garantiria resultado mais baixo em diversos dos indicadores. A cidade E foi criada com o mesmo critério das primeiras duas. Essa separação entre elas serviu para certificar que os algoritmos que estavam sendo usados realmente estavam agrupando com base nos indicadores, não com base em ordem na matriz ou algo diverso do que se queria evidenciar.

A cidade F foi criada manualmente, com 4 indicadores altos (sendo um deles o maior valor encontrado entre todas as cidades do grupo) e os outros baixos. A cidade G teve 2 indicadores altos, e o restante baixos. Os valores altos da cidade F e da cidade G não ocorrem nos mesmos indicadores. As cidades H, I e J foram criadas de forma a garantir que os valores fossem baixos, com distribuição uniforme de 0 a 30. A cidade K foi criada com os 5 primeiros indicadores da cidade J e os 5 últimos da cidade I; e a cidade L, a última, foi criada dividindo-se aproximadamente os valores da cidade J por dois. Esse aproveitamento de valores de uns e outros permitiu aferir-se a veracidade de algumas análises com base na matriz de correlação entre as cidades, já que se sabia o que esperar para os casos manipulados (correlação basicamente igual a 1 entre a cidade L e a cidade J).

A matriz de correlação entre os indicadores não revelou nada, uma vez que a relação entre os indicadores não foi pensada de antemão, sendo estes gerados quase que completamente aleatoriamente (não relacionados). Já a matriz de correlação entre as cidades permitiu que fossem conferidas algumas das relações preestabelecidas, principalmente aquelas entre as últimas cidades (I, J, K e L). A Tabela 4.3 abaixo mostra os valores da matriz calculada.

Tabela 4.3 Matriz de correlação entre cidades fictícias

	A	B	C	D	E	F	G	H	I	J	K	L
A	1.00											
B	0.03	1.00										
C	-0.02	-0.31	1.00									
D	-0.22	-0.39	0.54	1.00								
E	0.25	0.11	0.07	-0.25	1.00							
F	0.27	-0.10	0.47	0.32	-0.31	1.00						
G	-0.15	0.34	-0.02	0.16	-0.15	-0.31	1.00					
H	0.05	0.06	0.08	-0.03	0.27	0.21	-0.31	1.00				
I	0.47	0.41	-0.41	-0.06	0.13	0.02	0.03	0.43	1.00			
J	-0.42	-0.15	0.20	0.67	-0.06	-0.35	0.34	-0.13	0.12	1.00		
K	-0.11	0.50	-0.20	0.12	0.03	-0.11	0.60	0.11	0.61	0.43	1.00	
L	-0.44	-0.13	0.22	0.66	-0.05	-0.35	0.38	-0.13	0.10	1.00	0.46	1.00

Percebe-se uma correlação absolutamente positiva entre a cidade J e a cidade L (valor esperado, uma vez que uma é basicamente o dobro da outra em termos de indicadores), I e K, D e J, e D e L. Agrupando apenas estas 5 cidades, pode-se construir os histogramas para cada uma das cidades e suas correlações em um gráfico único, como mostrado na Figura 4.1. Este gráfico permite aferir como esta correlação alta se manifesta de forma geométrica através das nuvens de pontos geradas. Não se pode perceber distribuições de probabilidade muito claras nos histogramas devido à baixa dimensionalidade dos dados das cidades (apenas 10 características).

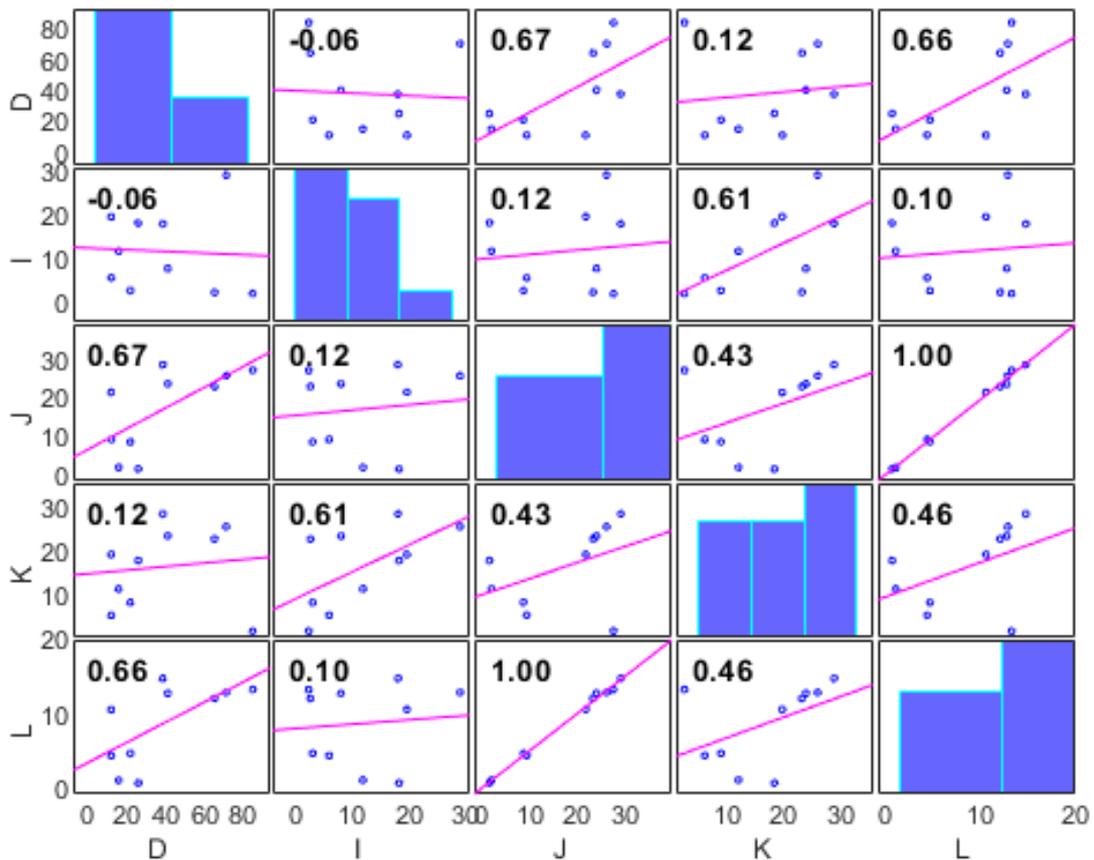


Figura 4.1 Matriz de correlação e histogramas de indicadores para as cidades D, I, J, K e L

Outra forma de se visualizar os dados simulados na Tabela 4.1 é o gráfico glifo (“*glyph plot*”). Este pode mostrar bem a distribuição final dos indicadores nas cidades visualmente. Trata-se de um grafo no qual cada uma das dimensões das observações, no caso atual, os indicadores, são apresentados como retas equidistantes saindo de um ponto, com a condição de que quanto maior o valor, maior o comprimento desta reta. Assim, permite-se visualizar com facilidade quais cidades têm os maiores indicadores. Caso o número de indicadores e de cidades seja razoavelmente baixo, esta visualização é extremamente fácil e intuitiva. Para

situações de mais indicadores ou mais cidades, esta representação acaba ficando menos útil. No caso atual de 10 indicadores, este gráfico pode substituir, qualitativamente, o uso de uma tabela para mostrar os dados finais usados. A Figura 4.2 mostra isso.

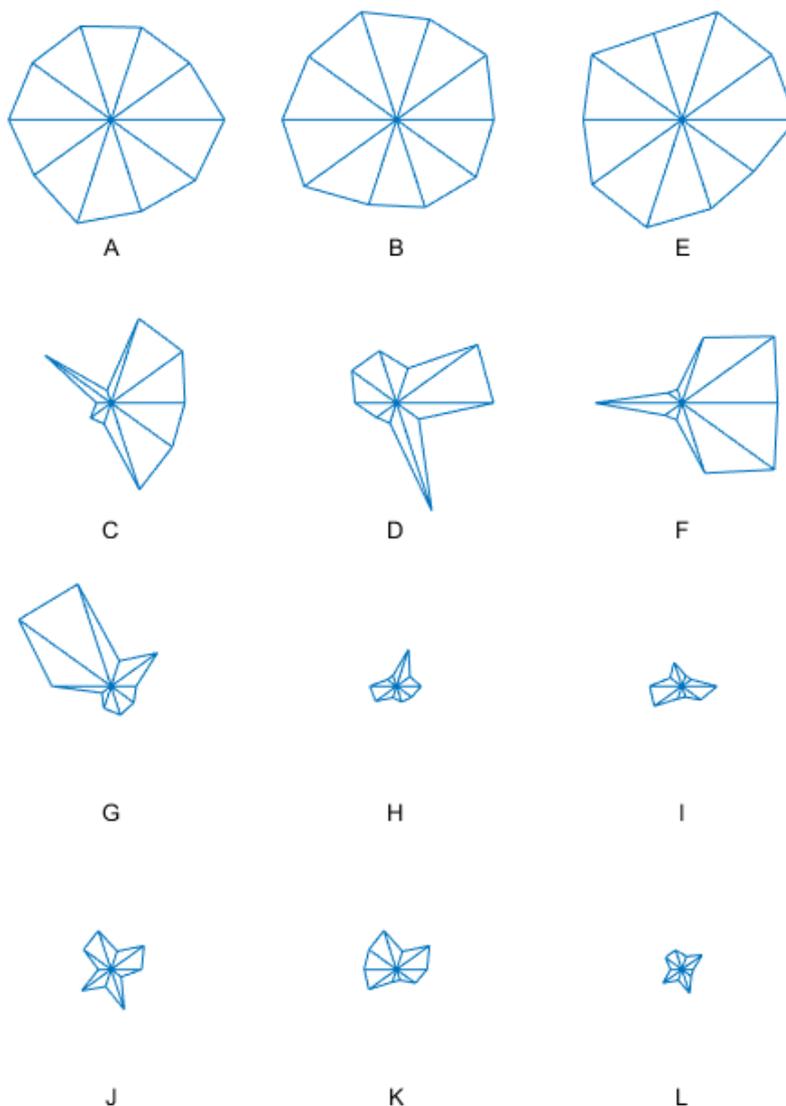


Figura 4.2 Distribuição dos indicadores nas cidades fictícias

Confirma-se, assim, o que já tinha sido mostrado na Tabela 4.1. As cidades A, B e E aqui postas em primeiro claramente apresentando os indicadores maiores. As cidades C, D, F e G têm perfil mais misto, com alguns maiores e outros menores (com ênfase ao que já havia sido dito de que os indicadores mais altos em F não são os mais altos em G). E, por fim, as cidades H, I, J, K e L sendo as com piores resultados. Nota-se também que a cidade L é claramente uma versão reduzida da cidade J, explicado pelo fato de que a segunda tem os valores de indicadores do dobro da primeira.

4.2. ANÁLISES ESTATÍSTICAS

4.2.1. Dendrograma

De posse dos dados fictícios, pôde-se iniciar as análises estatísticas. A primeira construção feita foi de um dendrograma. O algoritmo usado foi exatamente como descrito em 3.2.1. A imagem resultante foi a Figura 4.3 a seguir.

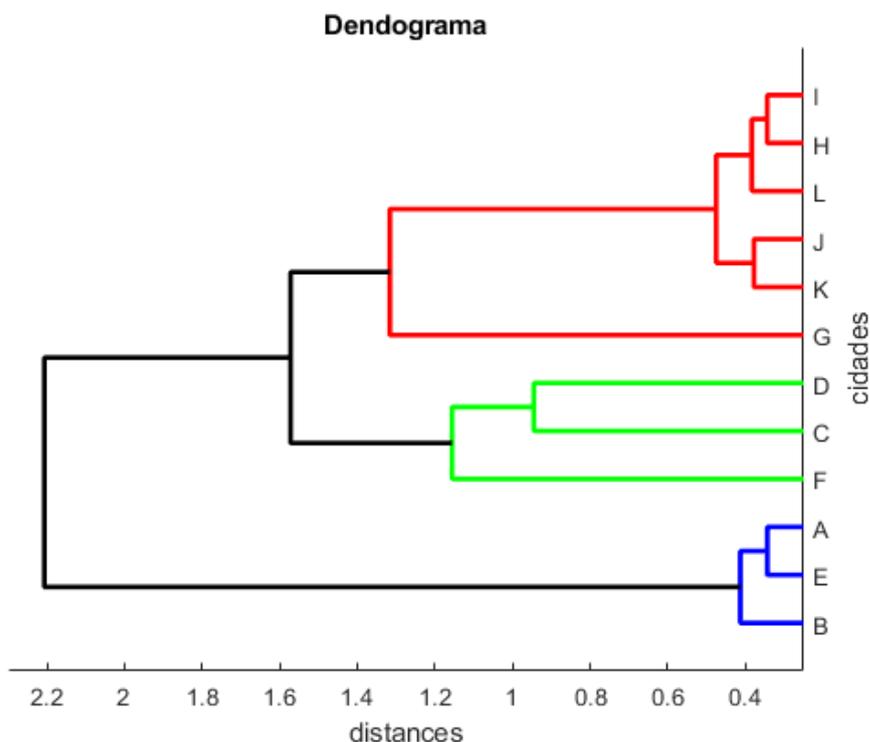


Figura 4.3 Dendrograma com dados fictícios de cidades

Aqui já se pôde perceber a formação de 3 grupos bem distintos de cidades, o azul de cidades mais bem graduadas (condizentes com aquelas mostradas na Figura 4.2), o verde com cidades intermediárias e o vermelho com aquelas que apresentaram os piores indicadores. Diferentemente do que se poderia assumir, a cidade G não ficou agrupada junto à C, D e F; mas acabou ficando praticamente em um grupo isolado. Levando isso em conta, os grupos criados por esta técnica foram sumarizados na Tabela 4.4 a seguir.

Tabela 4.4 Grupos identificados no dendrograma

Grupo	Cidades
1	A, B, E
2	C, D, F
3	G
4	H, I, L, J, K

4.2.2. Escalonamento multidimensional (MDS)

A próxima técnica utilizada foi o escalonamento, que também é capaz de criar grafos elucidantes. O algoritmo descrito em 3.2.2 foi utilizado para gerar uma nuvem de pontos em 3 dimensões, além das 3 projeções possíveis em 2-D. As imagens geradas foram as que compõe a Figura 4.4. Os grupos foram indicados manualmente pelas elipses vermelhas.

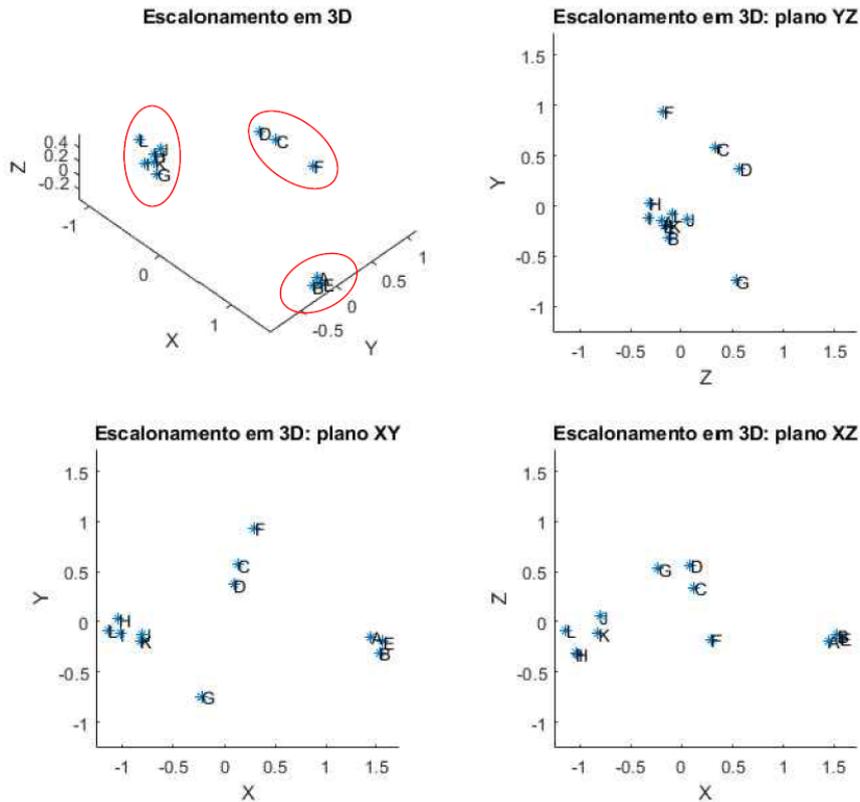


Figura 4.4 Escalonamento Multidimensional com grupos gerados

Vê-se principalmente nas projeções XY e XZ a formação de 3 colunas de pontos bem distintas: mais à direita, os 3 pontos com maiores indicadores (A, B e E), no centro, C e D, G e F; e todo o restante mais à esquerda. Esta disposição, ao contrário do encontrado na Figura 4.3, agrupou desde já G às cidades intermediárias, como seria intuitivo, mas ainda a mantendo afastada.

A projeção YZ não pareceu indicar nada além da similaridade entre os grupos, pois juntou A, B e E, muito similares por serem todas muito boas, às cidades H, I, J, K e L, também similares por serem muito ruins.

Assim, notou-se que não são suficientes as classificações com base nas projeções, sendo o resultado obtido para os dados em três dimensões mais completo, além de já gerar os resultados basicamente agrupados. As projeções ajudam na visualização dos grupos.

Os agrupamentos finais criados com base no escalonamento multidimensional são aqueles mostrados na Tabela 4.5, que não diferem dos mostrados na Tabela 4.4.

Tabela 4.5 Grupos identificados no escalonamento multidimensional

Grupo	Cidades
1	A, B, E
2	C, D, F
3	G
4	H, I, L, J, K

4.2.3. Análise de componentes principais (PCA)

Por fim, a análise de componente principal permitiu notar tanto a formação de grupos, como no escalonamento, quanto a relação entre as dimensões analisadas, os indicadores. No caso das cidades fictícias, as relações não são ainda de muita valia além da comprovação do método, mas no caso dos dados reais com cidades brasileiras, espera-se notar padrões de comportamento nas relações entre os indicadores, como aqueles encontrados nas análises de Torres-Salinas, Robinson-García, *et al.* (2013). Um exemplo plausível para o caso estudado seria o de que cidades com baixos índices de governança também apresentariam baixos índices de infraestrutura de transporte.

Para a escolha do número de componentes principais a serem analisadas, construiu-se o *scree plot* mostrado na Figura 4.5:

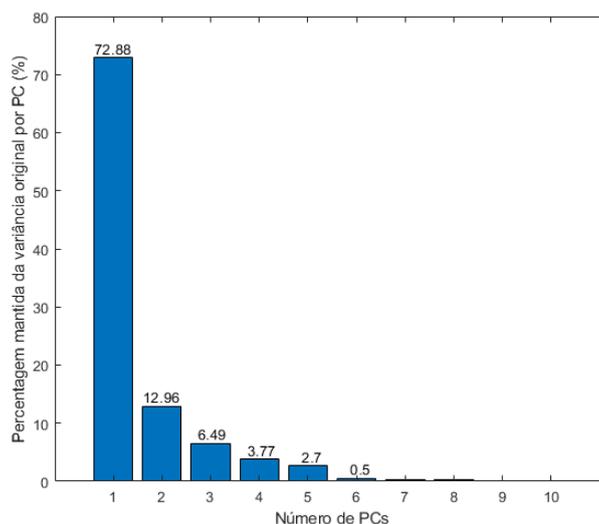


Figura 4.5 *Scree Plot* para dados de indicadores de cidades inteligentes

Com base na figura, vê-se que, com duas componentes principais, tem-se 85.84% da variância original. Uma dimensão a mais, incluindo a terceira componente, faz com que se

atinge 92.32%. Acréscimos da quarta e quinta componentes principais são modestos, mas ainda significativos; e a partir da sexta, não se fazem mais nem um pouco necessários. A visualização com mais três dimensões é dificultada. Dessa maneira, escolheu montar o *biplot* para duas e para três dimensões apenas.

A Figura 4.6 mostra o gráfico montado com as três primeiras componentes principais, já com os grupos formados indicados.

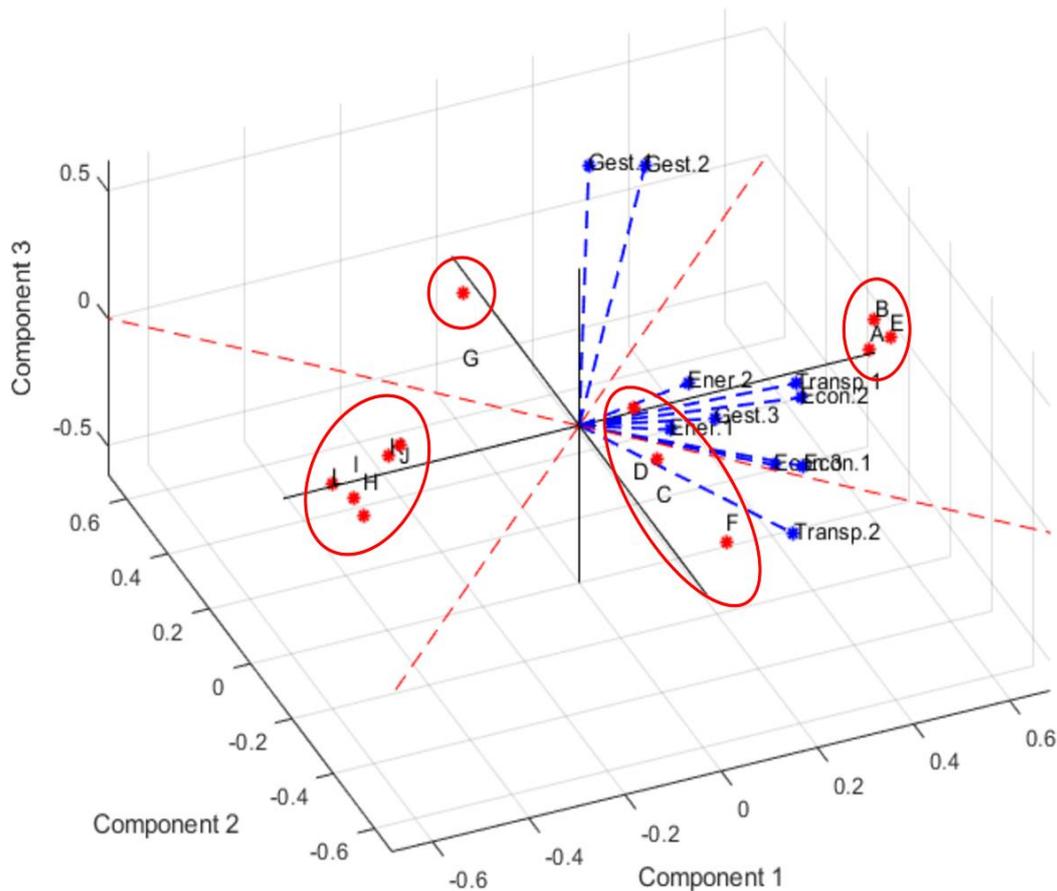


Figura 4.6 *Biplot* de Análise de Componente Principal 3D com grupos gerados

Percebe-se que uma das grandes dificuldades geradas pela visualização de *biplots* em três dimensões é que, embora a formação dos grupos se dê de forma clara, a relação entre as diversas características é dificultada. Uma forma de fazer com que estas relações se fizessem mais claras seria por meio das projeções, como foi feito para o MDS na Figura 4.4.

E a Figura 4.7 mostra o gráfico montado conforme o algoritmo descrito em 3.2.3, para duas dimensões, já com os grupos indicados.

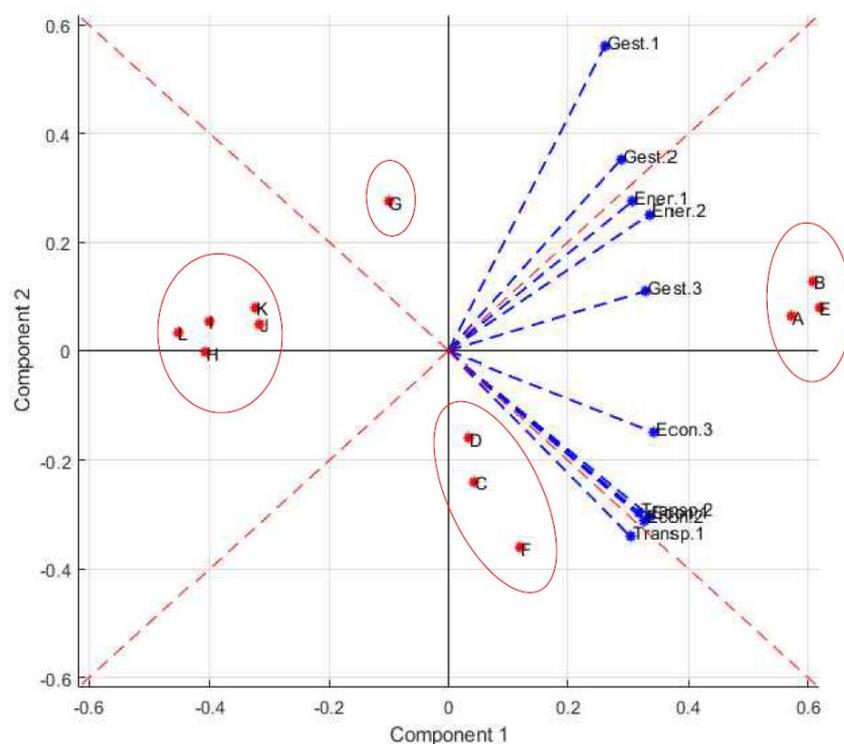


Figura 4.7 *Biplot* de Análise de Componente Principal 2D com grupos gerados

Embora o gráfico com as três primeiras componentes principais traga mais informação que o reduzido para duas (de 92.32% para 85.84%), os grupos continuam claramente formados mesmo na versão mais simples. Dessa maneira, a análise em duas dimensões, por ser mais simples, é mais útil.

Notou-se que o indicador Gest. 1 (Ind. 10.1) não tem quase relação alguma com Transp. 1 (Ind. 19.4), e este tem uma relação altamente positiva com Transp. 2 (Ind. 19.5), Econ. 1 (Ind. 5.2) e Econ. 2 (Ind. 5.2), que estão amalgamados perto deste primeiro. Como já dito, estas conclusões não são extrapoláveis aos casos reais, já que os dados usados de base foram completamente fictícios. Foram geradas quatro claras nuvens de pontos, agrupamentos estes que foram iguais aos mostrados para o dendrograma e que estão sumarizados na Tabela 4.6.

Tabela 4.6 Grupos identificados no PCA em 3D e em 2D

Grupo	Cidades
1	A, B, E
2	C, D, F
3	G
4	H, I, J, K

5. ESTUDO DE CASO: APLICAÇÃO A DADOS REAIS

5.1. COLETA DOS DADOS

Após demonstração da metodologia de agrupamento a partir de dados genéricos sintéticos, a próxima etapa foi a de aplicação a dados reais. Visando cumprir os objetivos definidos deste trabalho, então, buscou-se dados brutos ou já trabalhados referentes aos indicadores constantes nas normas ISO 37120 (2014) e ISO 37122 (2019). Em um primeiro momento, deu-se ênfase apenas ao cenário nacional, mas rapidamente notou-se uma escassez geral de dados que não sejam os colhidos por pesquisas nacionais abrangentes ou sistemas integrados como o DATASUS, do Ministério da Saúde, agrupados em grande parte no portal Cidades da página do IBGE. Dados básicos nacionais, estaduais e municipais constam nesta plataforma: pirâmide etária, taxa de mortalidade, domicílios com abastecimento de água, supridos por coleta de esgoto, acesso à internet, entre outros; todos com suas fontes referenciadas transparentemente (IBGE, 2017). Dessa maneira, a vasta maioria dos dados que compunham os indicadores constantes na normativa de 2019 não puderam ser encontrados já disponíveis. Ao leitor, convida-se a conferir novamente a Tabela 4.2 que dispõe o título de alguns indicadores ilustrativos usados na etapa anterior deste trabalho, constantes na ISO 37122: perceber-se-á a complexidade dos dados exigidos.

Em cidades estrangeiras também se notou uma escassez de dados referentes a esta Norma, mesmo em países de maior acumulação de capital como os componentes da Europa Ocidental e da América do Norte. Sendo assim, optou-se por não ser utilizada, para o estudo dos agrupamentos reais, a normativa mais recente.

A realidade nacional do monitoramento de alguns indicadores da ISO 37120 é mais promissora, seja por se tratar de um documento mais antigo e, portanto, mais consolidado, ou por abranger indicadores mais simples, que já têm seus dados coletados pelas mesmas pesquisas ou plataformas nacionais supracitadas. Entretanto, nenhuma cidade brasileira, na data de pesquisa, tinha a certificação ou mesmo estava em vias avançadas de conseguí-la. Brasília, entretanto, já havia demonstrado interesse em obtê-la por meio da criação do portal Brasília/DF em Dados (CODEPLAN, 2019). A página explicitamente assinala que a intenção futura é atingir os níveis de certificação. Os dados tinham periodicidade anual.

Nesta mesma plataforma, há uma ferramenta que faz possível comparar os valores dos indicadores obtidos por Brasília com aqueles de outras cidades no mundo, não apenas da Europa e Estados Unidos, mas também da América Latina, África e Ásia. Por este motivo, esta se configurou como a principal fonte de dados para as análises estatísticas reais.

Como não havia nenhuma outra cidade brasileira com esta facilidade, buscou-se em trabalhos acadêmicos os dados para quaisquer outras cidades. Fortuitamente, o cálculo de alguns dos indicadores da cidade do Rio de Janeiro foram encontrados no projeto de graduação de Elisa de Almeida Couto (2018). O indicador 18.1, não constante na base de dados apresentada por ela, foi estimado pelo autor do presente trabalho. Esta estimativa teve como base:

- As definições de transporte de alta capacidade, contidas na ISO 37120 (ISO, 2014);
- Os dados para Brasília do portal Brasília/DF em Dados (CODEPLAN, 2019);
- Dados populacionais e de malha atualizados da cidade do Rio de Janeiro, encontrados em sites de notícia.

Um grande entrave a ser superado foi: diversas cidades tinham dados incompletos (nem todos os indicadores constantes na plataforma da CODEPLAN ou em outras fontes). Como já observado previamente neste trabalho, os dados de entrada para os algoritmos devem ser completos *a priori*. Isso significa que os indicadores usados e cidades constantes deveriam ter todos os dados, não podendo haver cidades com lacunas. Por isso, seguiu-se os seguintes critérios para a escolha das cidades e dos indicadores:

- Tentar preservar maior número de cidades e indicadores, com indicadores de diversas áreas da Norma;
- Tentar manter um rol de cidades amplo, se possível ao menos uma de cada continente e ao menos um par de um só país e de realidades similares, para verificar o agrupamento final.

Realizadas as operações de exclusão de alguns indicadores e algumas cidades, obtiveram-se, por fim, as 14 cidades e 18 indicadores mostrados na Tabela 5.1 e

Tabela 5.2. Os números relativos na primeira facilitarão a visualização dos resultados futuramente.

Tabela 5.1 Indicadores selecionados

Área	Número relativo	Número do Indicador	Conteúdo
Economia	1	5.1	Taxa de desemprego da cidade
	2	5.3	Porcentagem da população abaixo da linha da pobreza

Área	Número relativo	Número do Indicador	Conteúdo
Educação	3	6.2	Porcentagem de estudantes com ensino primário completo
	4	6.3	Porcentagem de estudantes com ensino secundário completo
Finanças	5	9.1	Taxa de endividamento (expansão do serviço da dívida como uma porcentagem da receita própria do município)
	6	9.3	Porcentagem da receita própria em função do total das receitas
Governança	7	11.1	Porcentagem de participação dos eleitores nas últimas eleições municipais em função do total de eleitores aptos votar
	8	11.2	Porcentagem de mulheres eleitas em função do número total de eleitos na gestão da cidade
Saúde	9	12.1	Expectativa média de vida
	10	12.2	Número de leitos hospitalares por 100.000 habitantes
Segurança	11	14.2	Número de homicídios por 100.000 habitantes
Resíduos Sólidos	12	16.1	Porcentagem da população urbana com coleta regular de lixo (domiciliar)
Telecomunicações e Inovação	13	17.1	Número de conexões de internet por 100.000 habitantes
Transporte	14	18.1	Quilômetros de sistema de transporte público de alta capacidade por 100.000 habitantes
Planejamento Urbano	15	19.2	Número de árvores plantadas anualmente por 100.000 habitantes

Área	Número relativo	Número do Indicador	Conteúdo
Esgotos	16	20.1	Porcentagem da população da cidade atendida por sistemas de coleta e afastamento de esgoto
Água e Saneamento	17	21.1	Porcentagem da população da cidade com serviço de abastecimento de água potável
	18	21.2	Porcentagem da população da cidade com acesso sustentável a uma fonte de água adequada para consumo

Tabela 5.2 Cidades selecionadas

Continente	Cidades
América	Boston (EUA), Brasília (BRA), Buenos Aires (ARG), Portland (EUA) e Rio de Janeiro (BRA)
Europa	Aalter (BEL), Barcelona (ESP), Cambridge (UK), Eindhoven (HOL), Heerlen (HOL) e Koprivnica (CRO)
África	Cape Town (RSA)
Ásia	Taipei (TWN)
Oceania	Brisbane (AUS)

Vê-se, assim, que os objetivos citados anteriormente foram atingidos: há pelo menos uma cidade de cada um dos cinco continentes, há cidades de um mesmo país (Heerlen e Koprivnica, Boston e Portland, Brasília e Rio de Janeiro) e conseguiu-se manter um leque razoável de indicadores, abrangendo vasta parcela da ISO 37120. Em todos os gráficos criados para a análise desse conjunto de dados, utilizou-se a numeração relativa dos indicadores, conforme mostrado na segunda coluna da Tabela 5.1.

5.2. ANÁLISES ESTATÍSTICAS SEM REDUÇÃO DE INDICADORES

Da mesma forma como anteriormente com dados fictícios, os métodos selecionados de aprendizado estatístico foram aplicados aos dados, dessa vez, reais.

A primeira e mais direta conclusão obtida foi que diferentemente dos dados fictícios apresentados na Tabela 4.2, os dados reais não são tão dicotômicos. Isso quer dizer que a formação de grupos não se dá tão claramente quando se variando as técnicas de agrupamento. Toma-se, por exemplo, a expectativa de vida: o valor mínimo encontrado é 62,5 e o máximo de 83,4. Embora o leitor compreenda estes valores como radicalmente

diferentes, não os são da mesma forma que na tabela supracitada, com valores que em um mesmo indicador ora estavam na casa de uma dezena apenas, ora próximo à cem. A consequência prática disso é que as duas primeiras técnicas, em duas dimensões, o dendrograma e o MDS não foram capazes de criar grupos distintos e coerentes, como será mostrado.

5.2.1. Dendrograma

O dendrograma construído a partir de todos os 18 indicadores mostrados na Tabela 5.1 estão mostrados na Figura 5.1 a seguir:

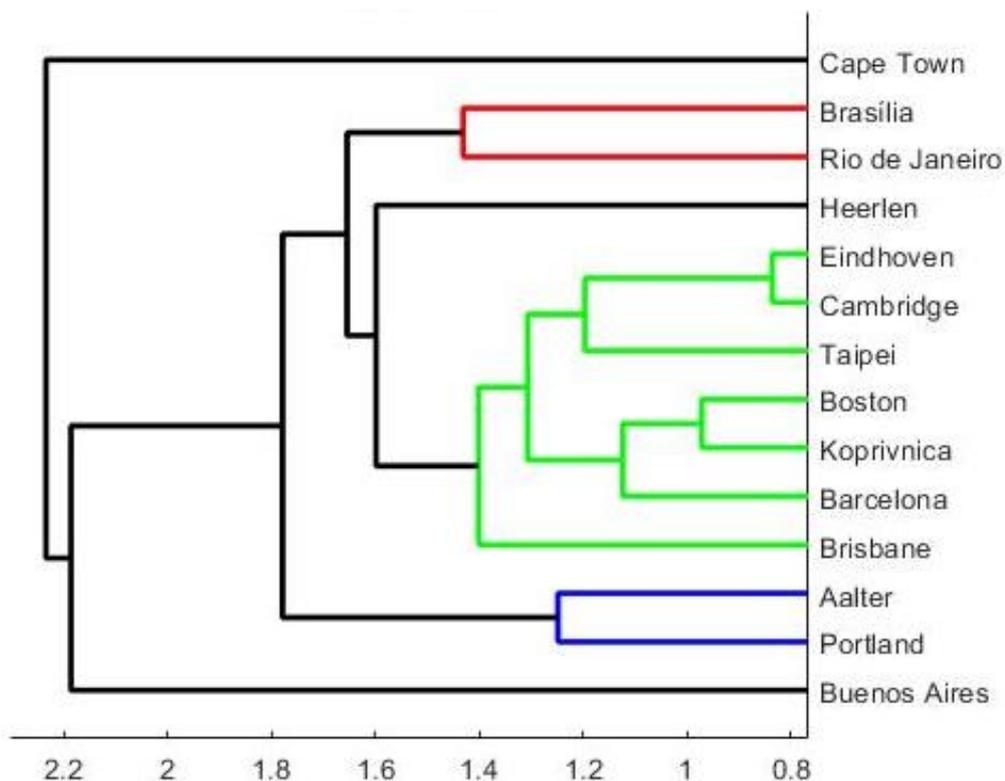


Figura 5.1 Dendrograma, 18 indicadores

Algumas incoerências já podem ser notadas: Buenos Aires está muito distante de outras cidades latino-americanas, como Brasília e Rio de Janeiro, além de também muito distante de Cape Town. Heerlen também está muito distante de outras cidades europeias. Não há grupos claramente formados para Buenos Aires, Cape Town ou Heerlen.

5.2.2. Escalonamento Multidimensional Bidimensional

O escalonamento multidimensional bidimensional foi outra técnica que acabou provando-se não muito útil. Os resultados para o cálculo com todos os 18 indicadores podem ser vistos na Figura 5.2 a seguir:

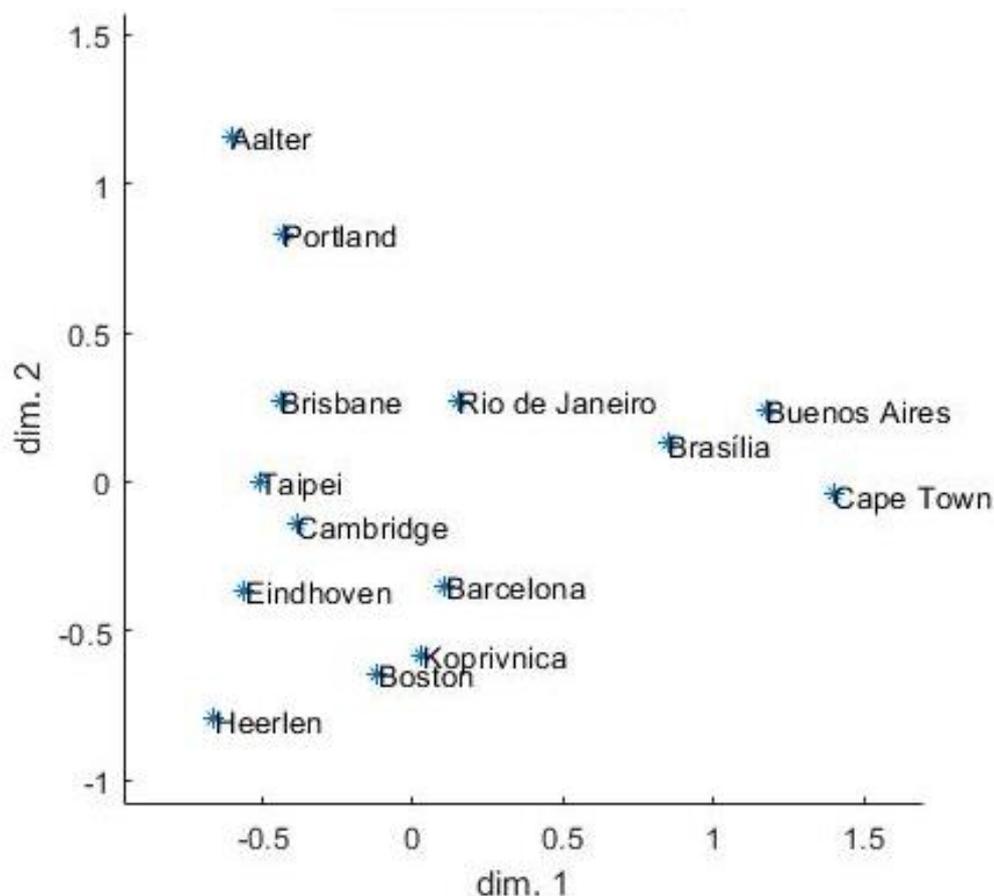


Figura 5.2 MDS 2D ou projeção XY, 18 indicadores

Aqui vê-se, por exemplo, que o Rio de Janeiro está mais próximo de Brisbane do que da outra cidade brasileira, Brasília, contrariando vorazmente o dendrograma mostrado na Figura 5.1. Não é possível, portanto, usar ambas as técnicas para formar grupos iguais ou mesmo muito similares, não havendo complementariedade entre elas.

Atribui-se estas incompatibilidades ao fato de que, com dados de diferenças mais sutis como os reais, estas técnicas bidimensionais acabam se tornando muito simplistas.

5.2.3. Escalonamento Multidimensional Tridimensional

Pelos motivos mencionados, olvidam-se os resultados anteriores, considerando-os ineptos ao estudo proposto. A adição de outra dimensão à técnica anterior faz com que surjam grupos mais claros. A imagem a seguir, unida à Figura 5.2, que é a projeção XY do gráfico tridimensional, permite ao leitor ter uma noção boa da verticalidade dos dados. Ou seja, enquanto Buenos Aires e Cape Town pareciam bem próximos em Figura 5.2, em três dimensões, na Figura 5.3, já se os vê bem afastados.

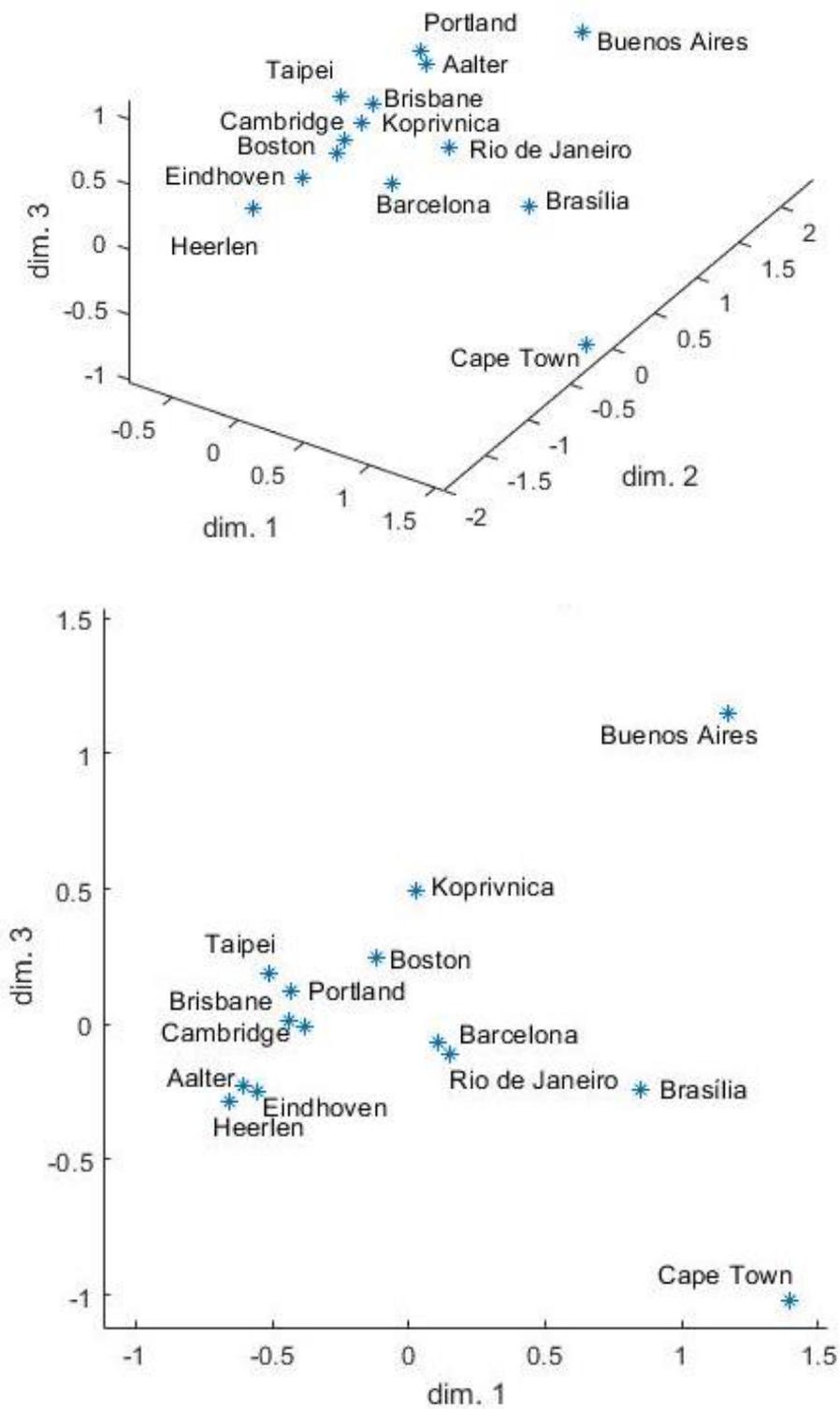


Figura 5.3 MDS 3D e projeção XZ, 18 indicadores

Como a tridimensionalidade é crucial, os grupos, embora formados claramente na visualização livre, são de difícil traçado e indicação bidimensional. Por isso, eles são

mostrados apenas em formato de tabela na e fica o leitor livre para reconstituí-los a partir das projeções mostradas anteriormente.

Tabela 5.3 Grupos formados a partir do MDS 3D

Grupos ¹	Cidades
I	Aalter, Portland
II	Rio de Janeiro, Brasília, Cape Town, Buenos Aires
III	Heerlen, Eindhoven
IV	Cambridge, Taipei, Brisbane
V	Boston, Barcelona, Koprivnica

¹Grupos não dispostos em ordem alguma de indicadores

Vê-se agora, diferentemente do dendrograma e do MDS 2D um agrupamento muito mais convincente: cidades similares estão agrupadas. Rio de Janeiro, Brasília e Buenos Aires, todas latino-americanas, ficam juntas. Taipei, cidade taiwanesa avançadíssima, agrupa-se a Cambridge e Brisbane, anglófonas de países ricos, da Europa e Oceania. Disso conclui-se que as cidades em desenvolvimento criaram juntas uma categoria enquanto as cidades desenvolvidas separaram-se em subgrupos. Nota-se, entretanto, a partir da Figura 5.3, que a distância entre estas é bem menor que entre aquelas.

5.2.4. Análise de Componente Principal Tridimensional

Como mostrado, as análises bidimensionais não foram suficientes. O *scree plot* da PCA evidencia mais ainda este fato, mostrando que a variância agregada na terceira dimensão ainda é longe de irrelevante.

Uma versão levemente alterada do *scree plot* pode ser ainda mais útil. Se ao invés de dispor os valores adquiridos de variância por nova componente principal, dispõem-se os valores de variância adquirida acumulada, vê-se com mais clareza a importância da confecção do gráfico em três dimensões. Esta versão acumulada está mostrada na Figura 5.4 a seguir:

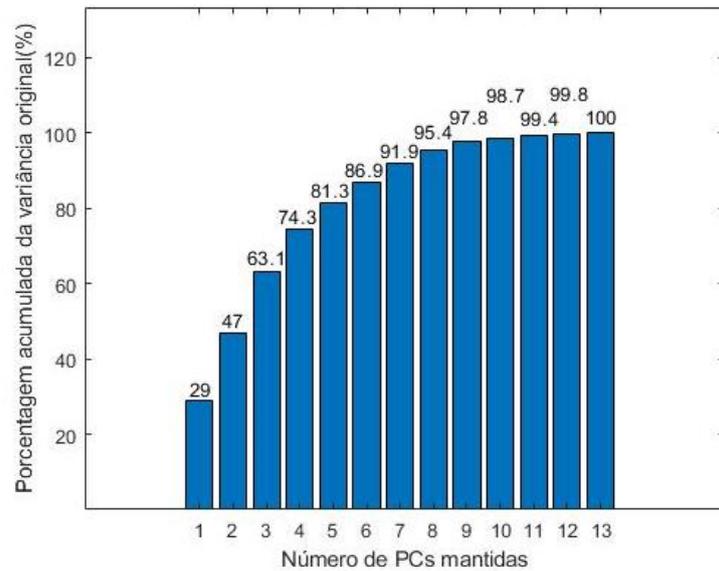


Figura 5.4 *Scree Plot* acumulado para 18 indicadores

Vê-se que apenas com duas dimensões nem metade da variância da amostra original está presente na análise PCA. Há, entretanto, um pulo de mais de 15% da segunda para terceira componente principal. Embora seja um valor abaixo daquele mostrado na Figura 4.5 para os exemplos fictícios, considera-se que seja suficiente para a formação dos grupos, como poderá ser visto. O *biplot* construído foi como mostrado a seguir:

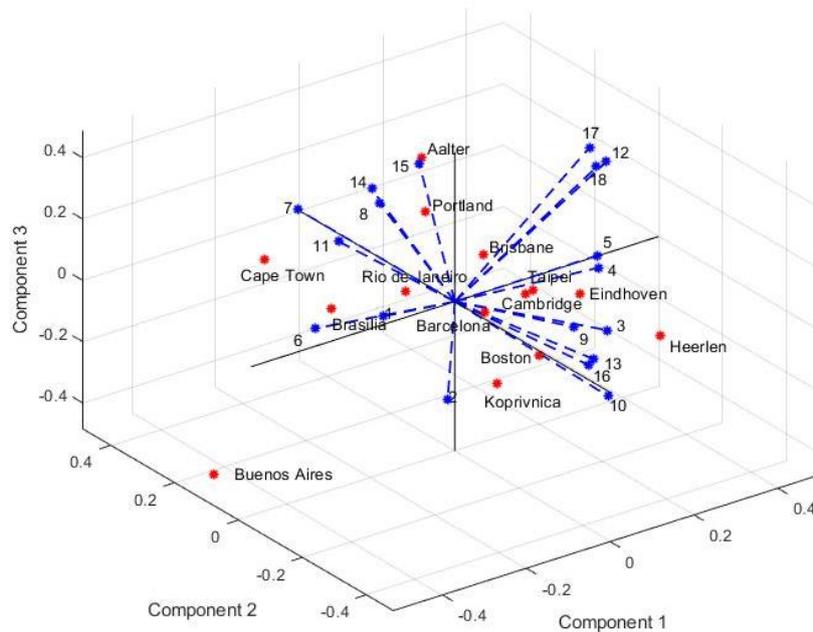


Figura 5.5 *Biplot* 3D com 18 indicadores

Os mesmos empecilhos encontrados na visualização tridimensional da Figura 5.3 são encontrados aqui. A partir da visualização livre do gráfico, montam-se os agrupamentos, mostrados na Tabela 5.4 a seguir:

Tabela 5.4 Grupos formados a partir do *biplot* tridimensional com 18 indicadores

Grupos ¹	Cidades
I	Aalter, Portland
II	Rio de Janeiro, Brasília, Cape Town, Buenos Aires
III	Heerlen, Eindhoven
IV	Cambridge, Taipei, Brisbane
V	Boston, Barcelona, Koprivnica

¹Grupos não dispostos em ordem alguma de indicadores

Vê-se que os grupos aqui formados são precisamente aqueles já mostrados na Tabela 5.3. Outra conclusão interessante que era esperada para o caso real e que não acontecia no caso fictício é um estudo esmiuçado das relações entre os indicadores, mostradas graficamente pelos vetores em azul. Estão muito relacionados 3 e 9 (ver índices relativos na Tabela 5.1), mostrando que a porcentagem dos estudantes com ensino primário completo é altamente ligada à expectativa média de vida de uma população, algo lógico e talvez esperado. Os indicadores 12, 17 e 18 também tem forte correlação: são abastecimento de água, água potável e coleta regular de lixo. Da mesma forma, espera-se que em cidades com melhores serviços de distribuição de água, haja melhores serviços de coleta de lixo; outro dado lógico e esperado. Uma relação não esperada, mas não menos interessante é entre 14 e 8, mostrando que em cidades onde há mais mulheres eleitas, nas rédeas das decisões públicas, há mais linhas de transporte público de alta capacidade. Talvez esta constatação indique um maior apreço feminino pelo bem-estar público; talvez que em países que chegam a um nível tal de acumulação de capital, tem-se o luxo de transporte público de alta densidade e da participação feminina competitiva em eleições, condicionando o tratamento humano do ser humano às veleidades do Capital. Outras relações que aparecem no *biplot* não são tão claras ou tão interessantes a ponto de serem comentadas.

5.3. ANÁLISES ESTATÍSTICAS COM REDUÇÃO DE INDICADORES

5.3.1. 13 Indicadores (5 a menos que o original)

Em se tratando de visualizar correlações, embora o *biplot* gerado pela análise de componentes principais permita uma visualização gráfica facilitada, talvez a forma mais consagrada e conhecida é a do cálculo da matriz de correlação. Na etapa anterior deste trabalho, foi calculada uma matriz de correlação entre as cidades fictícias estabelecidas, mostradas na Tabela 4.3. Para as análises que se pretende realizar agora, mais interessante seria a construção da matriz de correlação entre as variáveis, ou seja, os indicadores. A partir das relações reveladas por esta, poder-se-ia retirar alguns dos indicadores e, em tese, manter os mesmos agrupamentos. Isso ocorreria pois, se dois indicadores têm alta relação entre si, para o caso dos agrupamentos, a coleta dos dois é supérflua, sendo um apenas já capaz de distanciar as cidades entre si. As tabelas Tabela 5.5 e Tabela 5.6 a seguir mostram esta matriz de correlação para os indicadores, indicados pela numeração relativa constante na Tabela 5.1.

Tabela 5.5 Matriz de correlação entre indicadores reais (Parte 1)

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00											
2	0.48	1.00										
3	-0.74	-0.18	1.00									
4	-0.63	-0.18	0.85	1.00								
5	-0.48	-0.37	0.50	0.53	1.00							
6	-0.18	0.17	-0.06	-0.08	-0.54	1.00						
7	0.02	-0.47	-0.30	-0.14	-0.12	0.08	1.00					
8	-0.03	0.05	0.09	0.24	-0.30	0.25	0.08	1.00				
9	-0.66	-0.52	0.80	0.60	0.28	-0.08	-0.07	0.06	1.00			
10	-0.19	0.39	0.39	0.37	-0.06	-0.06	-0.47	-0.19	0.22	1.00		
11	0.64	0.27	-0.85	-0.64	-0.26	0.08	0.20	-0.18	-0.92	-0.29	1.00	
12	-0.18	-0.19	0.30	0.51	0.45	-0.32	-0.24	0.18	0.13	0.20	0.01	1.00
13	-0.06	0.00	0.19	0.15	0.48	-0.59	-0.17	-0.43	0.10	0.29	-0.21	0.06
14	-0.36	-0.16	0.19	0.21	0.14	0.13	0.28	0.72	0.07	-0.31	-0.17	0.13
15	-0.26	-0.36	0.08	0.35	0.10	0.11	0.56	0.31	0.10	-0.30	-0.04	0.31
16	0.28	0.21	0.03	0.01	-0.10	-0.22	-0.50	-0.20	0.16	0.24	-0.17	0.16
17	-0.13	-0.34	0.15	0.33	0.37	-0.38	-0.15	0.16	0.15	0.08	0.04	0.95
18	-0.09	-0.19	0.15	0.36	0.29	-0.38	-0.22	0.16	0.11	0.23	0.03	0.90

Tabela 5.6 Matriz de correlação entre indicadores reais (Parte 2)

	13	14	15	16	17	18
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13	1.00					
14	-0.14	1.00				
15	-0.05	0.35	1.00			
16	0.30	-0.58	0.00	1.00		
17	0.05	0.12	0.32	0.20	1.00	
18	-0.02	0.03	0.23	0.17	0.93	1.00

Nas tabelas Tabela 5.5 e Tabela 5.6 mostradas, foram assinalados os valores de correlação considerados muito altos: acima de 0.80 ou abaixo de -0.80. Estes valores mostram que os indicadores têm quase o mesmo peso na definição dos grupos (vetores quase sobrepostos no *biplot*) ou são absolutamente opostos (vetores de mesma direção, mas sentidos diferentes no *biplot*). Os pares de alta correlação formados foram, assim: (4,3); (9,3); (11,3); (11,9); (17,12); (18,12) e (18,17). Com base nesses pares encontrados, o maior número de indicadores considerados supérfluos que podem ser retirados de forma a tentar manter os mesmos agrupamentos é cinco: 4, 9, 11, 17 e 18. Dessa forma, 12 cumpriria o papel de 17 e 18; e 3 representaria 4, 9 e 11. Com base apenas na análise da matriz de correlação, foi-se capaz de eliminar cinco indicadores, 28% dos indicadores iniciais.

Para verificar a formação dos grupos, optou-se por usar somente a técnica PCA, por se tratar da mais completa e mais relevante, como já mostrado. Primeiramente, o *scree plot* para verificar se houve muita diferença da variância total com três dimensões daquela encontrada

na Figura 5.4. Reitera-se que o gráfico construído é o acumulado, mostrado na Figura 5.6 abaixo:

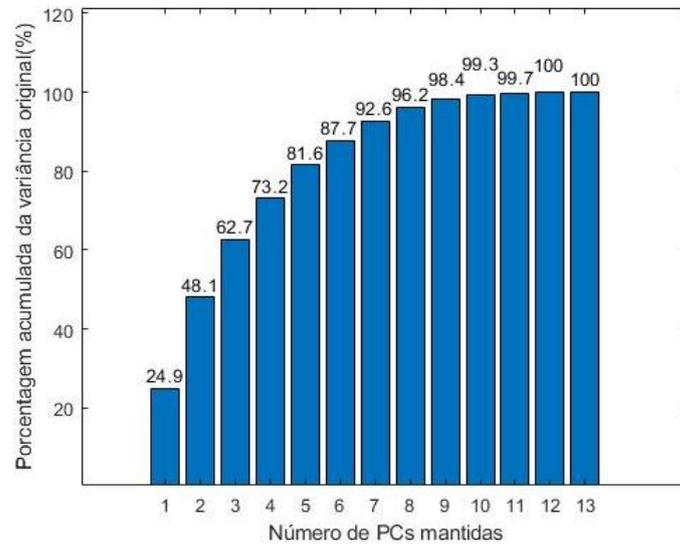


Figura 5.6 *Scree Plot* acumulado para 13 indicadores

Observa-se, assim, uma perda de apenas 0.4% com três componentes principais mantidas em relação ao valor encontrado para os 18 indicadores. Considera-se essa diferença irrelevante. O *biplot* construído a partir destes indicadores para as mesmas cidades é:

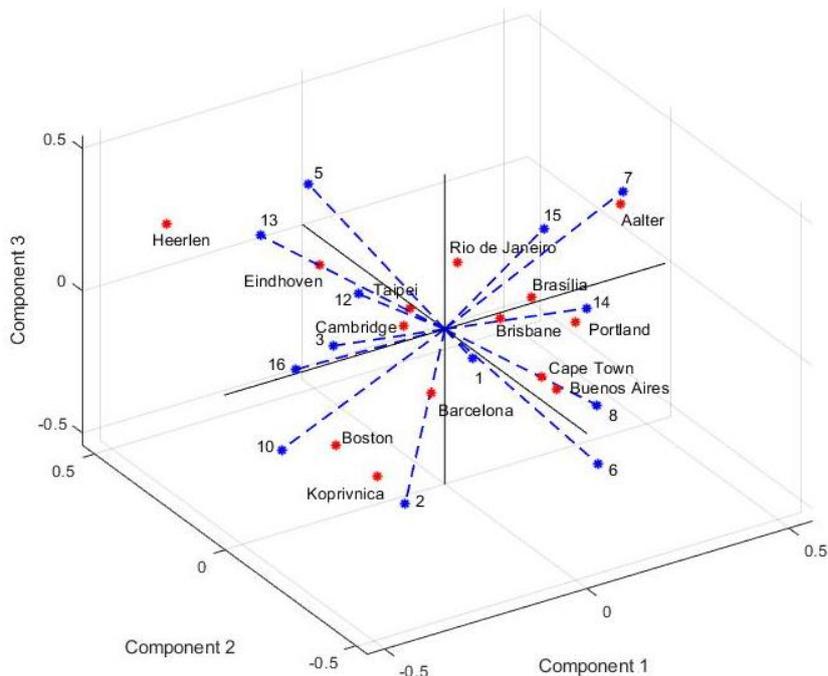


Figura 5.7 *Biplot* 3D com 13 indicadores

Neste caso, diferentemente do mostrado na Figura 5.5, a visualização é mais clara devido ao menor número de vetores azuis que representam os indicadores. Pode-se notar também

alguns dos grupos assinalados já, enquanto outros se mostram de mais difícil visualização pela via isométrica estática. O caso das cidades em desenvolvimento (Rio de Janeiro, Brasília, Buenos Aires e Cape Town) é um que pode enganar: embora na Figura 5.7 elas pareçam muito próximas a cidades desenvolvidas como Brisbane e Portland, no eixo Z estas cidades estão muito distantes. Enquanto as cidades desenvolvidas estão na porção negativa do eixo, todas as outras se encontram na porção positiva. Assim, os grupos mantêm-se inalterados. São eles os da Tabela 5.7 a seguir:

Tabela 5.7 Grupos formados a partir do *biplot* tridimensional com 13 indicadores

Grupos ¹	Cidades
I	Aalter, Portland
II	Rio de Janeiro, Brasília, Cape Town, Buenos Aires
III	Heerlen, Eindhoven
IV	Cambridge, Taipei, Brisbane
V	Boston, Barcelona, Koprivnica

¹Grupos não dispostos em ordem alguma de indicadores

5.3.2. 12 Indicadores (6 a menos que o original)

Uma forma plausível de se reduzir ainda mais o número de indicadores é a partir do próprio *biplot*. Como já explicado, os vetores em azul indicam a correlação entre as variáveis, com vetores na mesma direção muito relacionados positivamente (correlação próxima de 1) e vetores em direções opostas muito relacionados negativamente (correlação próxima de -1).

Pode-se pensar o *biplot* bidimensional como sendo a sombra ou projeção do gráfico tridimensional associado. Dessa forma, *per analogiam*, pode-se entender o *biplot* em três dimensões como sendo a sombra tridimensional do gráfico “completo”, ou seja, em treze dimensões nos casos vistos até aqui (número de componentes principais). Lembra-se que a representação com três eixos manteve em torno de 60% da variância total, que é *grosso modo* a proporção de informação presente em relação ao conjunto de dados original. É por esse

motivo que mesmo depois de retirados todos os indicadores correlacionados, ainda há vetores azuis pertos do paralelismo.

Na Figura 5.7, pode-se observar que os vetores 10 e 7 aparentam estar em direções opostas, além de 6 e 13. De fato, o ângulo formado entre estes, que pode ser facilmente calculado analiticamente, é muito próximo de 180° ($173,5^{\circ}$ no primeiro caso e $163,3^{\circ}$ no segundo). Ou seja, pode-se pensar em retirá-los para fins de economia.

No primeiro caso, retirou-se apenas o indicador 10, por ser o ângulo formado entre seu correlato muito próximo de 180° , o que permitiria ter certeza de que fazer essa retirada é viável e manteria os mesmos agrupamentos. A escolha do 10 em detrimento do 7 foi devido ao módulo do vetor dele, que era levemente menor. Em questão de custo de obtenção, ambos são dados disponíveis na internet. Manteve-se assim um grupo de 12 indicadores, ou 33,3% a menos que o original. O *scree plot* com 12 indicadores, acumulado, está mostrado na Figura 5.8 a seguir:

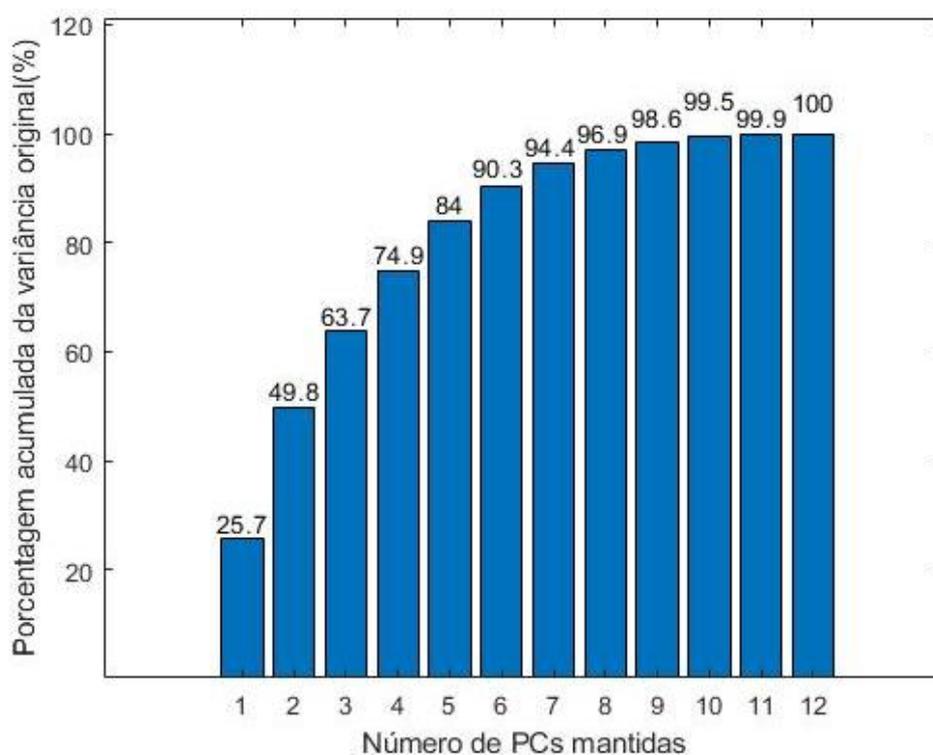


Figura 5.8 *Scree Plot* acumulado para 12 indicadores

Vê-se que pouco se altera na variância total obtida com cada dimensão conforme se reduz o número de indicadores, apenas que o número de componentes principais máximo caiu de 13 para 12 devido à queda no número de indicadores. Isso ocorre pois em uma amostra onde o número de observações não é o limitante, o limitante para as componentes

passa a ser o número de variáveis. A partir do momento que se passou a ter 12 variáveis, o número de PCs caiu para também 12. O *biplot* com estes dados reduzidos foi construído como mostrado na Figura 5.9:

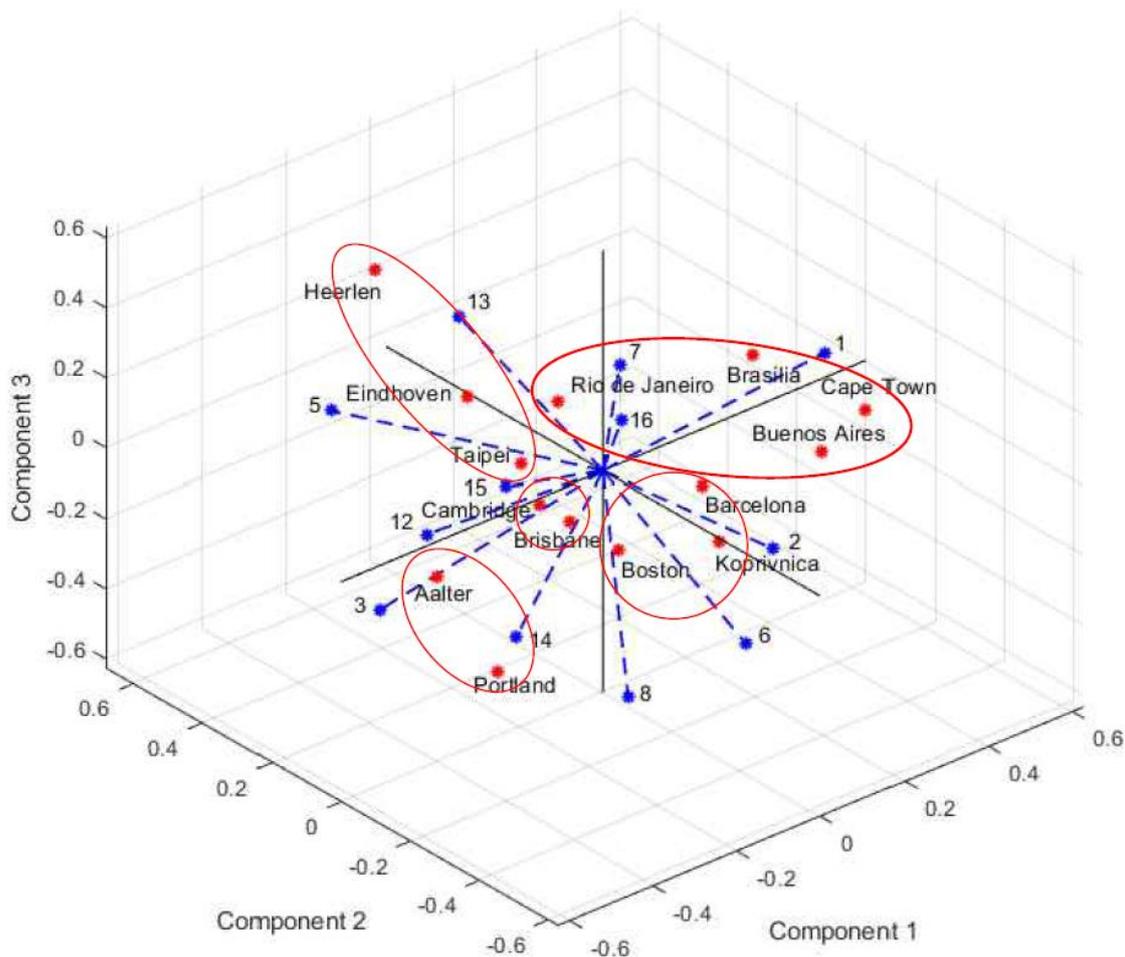


Figura 5.9 *Biplot* 3D com 12 indicadores

Diferentemente do que vem acontecendo neste trabalho até então, desta vez os pontos em três dimensões ficaram de tal forma organizados que pôde-se destacar os grupos na própria imagem. O leitor pode assim comparar os grupos mostrados na imagem àqueles dispostos na Tabela 5.7 e Tabela 5.4 e ver que estes são iguais. Ou seja, mostrou-se possível a redução do número de indicadores sem qualquer prejuízo na formação dos grupos.

5.3.3. 11 Indicadores (7 a menos que o original)

Nessa mesma Figura 5.9, pode-se notar o par de vetores 6 e 13, já citados anteriormente, novamente aparecendo no mesmo sentido, mas direções opostas. Para verificar se os 163,3 graus entre eles podem ser considerados próximos o suficiente de 180 e para tentar reduzir ainda mais o número de indicadores, retirou-se um deste par. Como o indicador 6

(porcentagem da receita própria em função do total das receitas) é um indicador mais complexo que o 13, que versa sobre acessos a internet per capita, dado já presente no portal Cidades do IBGE; foi ele deixado de lado. O grupo de indicadores passou a ser, dessa forma, 39% menor que o original.

O *scree plot* acumulado para esta nova amostra com 11 indicadores foi plotado como na Figura 5.10 a seguir:

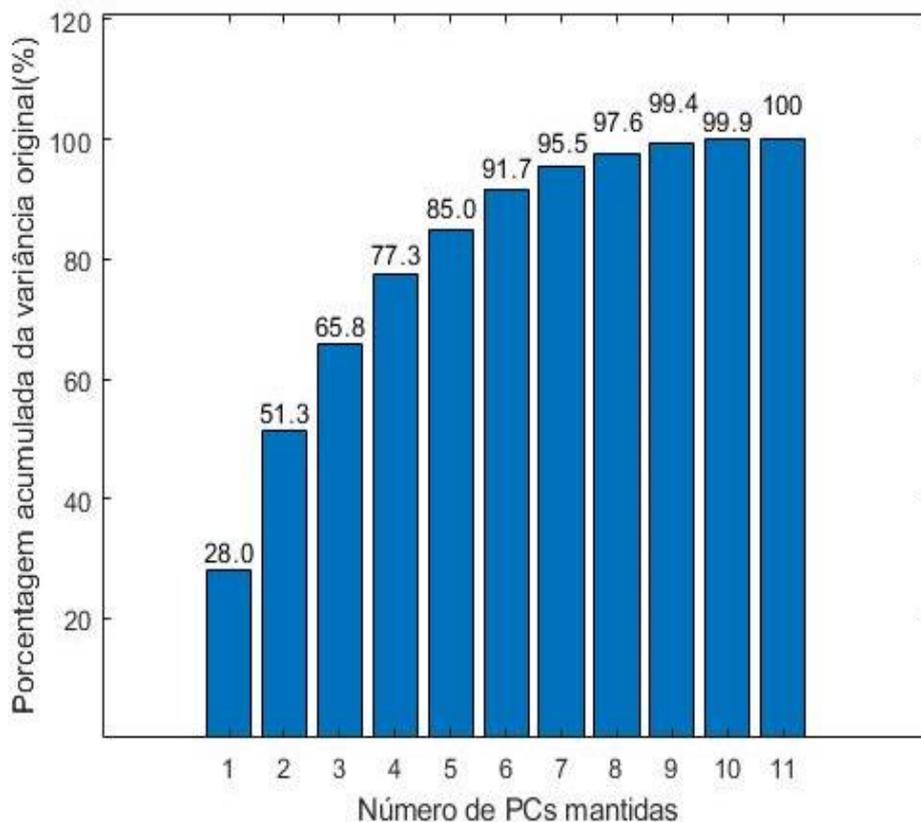


Figura 5.10 *Scree Plot* acumulado para 12 indicadores

Neste caso o número de componentes principais totais também caiu de 12 para 11. O formato e valores constantes neste gráfico, entretanto, são tão similares aos mostrados para os outros casos que não serve para tirar conclusões relevantes; o mesmo patamar de 60% da variância original em três dimensões continua sendo atingido, com um leve salto neste último caso atribuído, voltando à analogia da sombra, ao fato de que a o “objeto real” tem menos dimensões neste exemplo, sendo melhor representado nas mesmas três dimensões.

O *biplot* para o caso de 11 indicadores é o mostrado na Figura 5.11 a seguir:

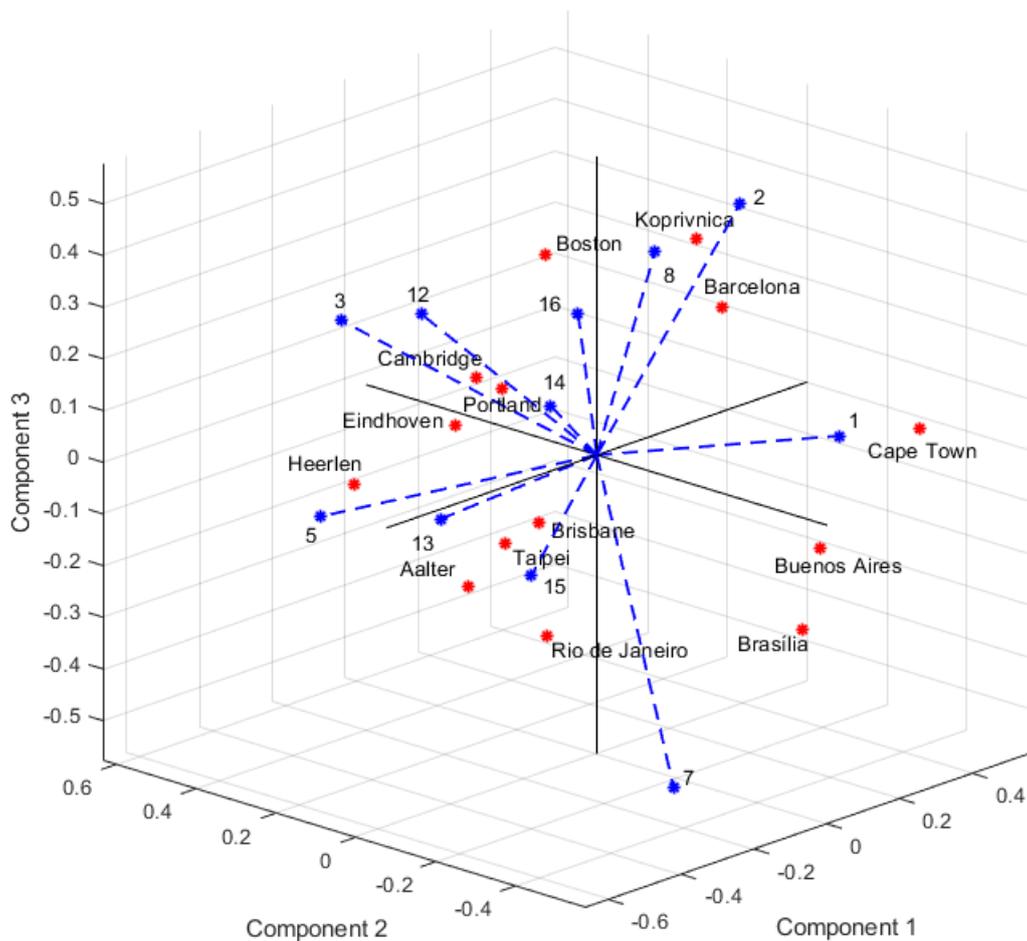


Figura 5.11 *Biplot* 3D com 11 indicadores

A primeira coisa que salta aos olhos é a multidirecionalidade dos vetores, agora que nenhum deles está nem próximo do estar mesmo sentido. Considera-se, assim, que se atingiu o mínimo absoluto de variáveis para esta amostra seguindo o método de retirada via *biplot*. No tocante aos grupos, vê-se que alguns mantêm-se os mesmos, como o Grupo II e V da Tabela 5.7. Na visualização livre, também se pode perceber que o Grupo I ainda é mantido. Os grupos III e IV, no entanto, foram alterados. Cambridge, como se pode ver na Figura 5.11, aproxima-se muito de Eindhoven. Em três dimensões, considerar-se-ia que, de fato, aquela comporia um grupo com esta cidade e Heerlen. Dessa forma, considera-se já não satisfatória a redução, uma vez que alterou os grupos.

Por esta inconsistência na formação dos agrupamentos, considera-se que este método de retirada de indicadores (pela análise do *biplot* tridimensional) não é confiável. Os fatores que levaram a esta conclusão foram: i) a tolerância usada para o ângulo formado entre os vetores que se pretende suprimir, devendo, talvez, ser mais estreita, ii) a tolerância para uso da técnica de retirada pelo *biplot* da variância total mantida nas três dimensões analisadas talvez

também deva ser mais estreita. Em todo caso, para garantia do resultado do agrupamento, as retiradas só deverão ser feitas pela matriz de correlação entre as variáveis.

6. CONCLUSÕES

As técnicas propostas neste trabalho não são recentes, nem mesmo as aplicações aqui sugeridas. Notou-se, no entanto, a falta de um trabalho que se propunha, além de outras coisas, a estudar as diferenças entre elas e sua capacidade específica de gerar agrupamentos, ainda mais para o caso de cidades inteligentes e sustentáveis. Os agrupamentos são interessantes de serem estudados pois poderiam permitir um melhor compartilhamento de experiências entre cidades que se encaixam em um mesmo grupo, ou mesmo de cidades-tutoras ajudando cidades-aspirantes.

Infelizmente as análises com o corpo normativo mais recente, a ISO 37122 (2019), não foram possíveis devido ao verdadeiro deserto de dados existente. A realidade para a norma anterior desta mesma família, no entanto, se mostrou muito mais promissora e a plataforma da Companhia de Planejamento do Distrito Federal, Brasília/DF em Dados, já possuía os dados de Brasília, como também de diversas outras cidades internacionais. O cenário brasileiro, embora promissor, provou-se escasso para todos os indicadores que fugiam das coletas sazonais realizadas pelo IBGE ou pelos sistemas agregados aos Ministérios da Economia, Educação e Saúde.

Visando pôr à prova real os métodos primeiramente testados em valores fictícios, usaram-se os dados desta plataforma e alguns outros de outras fontes (em especial os relacionados à cidade do Rio de Janeiro) relativos aos indicadores da norma mais antiga. Buscou-se além de obter um rol amplo de cidades, obter também uma análise por completo da norma, com uso de indicadores de diversas porções do documento. Também se tentou fugir ao máximo dos indicadores mais difundidos, como expectativa de vida ou renda per capita.

Os dados reais mostraram-se não tão dicotômicos quanto aqueles simulados anteriormente, fazendo com que as diferenças entre os grupos se reduzissem. Dessa maneira, o uso das técnicas bidimensionais (dendrograma, escalonamento multidimensional bidimensional) não foi o suficiente para a delimitação dos grupos. As técnicas mais elaboradas e tridimensionais (MDS 3D e análise de componente principal), entretanto, se provaram mais do que capazes de exacerbar as diferenças existentes e permitir a formação dos grupos.

Um dos grandes entraves à geração de análises com indicadores é o custo da coleta de dados ou de seu tratamento. Por mais que um indicador possa vir a contribuir para a construção de uma informação valiosa, caso ele seja muito difícil de ser constituído, poucos

serão os dispostos a usá-lo. Por esse motivo, quando se tratando de agrupamentos, seria interessante existir uma forma de se reduzir a coleta de dados, em particular dos dados mais custosos. Na última parte deste trabalho, dedicou-se especialmente a este caso, propondo uma solução de relativa facilidade de implementação. A partir da matriz de correlação, pôde-se resumir o conjunto inicial de indicadores a apenas 72,2% do seu tamanho original ainda mantendo os agrupamentos originais com bastante clareza. Não foram todas as escolhas de qual indicador seria mantido ou não que tiveram muito espaço para serem feitas levando em conta o quesito custo de indicador, mas em estudos mais robustos, com maior número de indicadores, isso com certeza deverá ser levado em conta. Somente para ilustração, caso fosse utilizada esta metodologia para o orçamento do censo brasileiro de 2 bilhões de reais, a economia seria estimada em 555 milhões. Outra maneira ainda de se trabalhar seria a partir de índices que resumissem melhor as múltiplas realidades das cidades, para então gerar os agrupamentos.

Ademais, provou-se aqui que a verificação da correlação entre variáveis e eliminação daquelas redundantes (por terem correlação próxima à absoluta, positiva e negativamente) é um método razoável para economia computacional e financeira na geração de agrupamentos. Para esta economia ser de fato efetivada, dever-se-ia construir um corpo de conhecimento que traçaria as correlações médias entre os mais diversos indicadores que se deseja estudar. De posse destes dados, escolher-se ia o menor número possível de indicadores para serem levados em conta, também pesando na balança o custo de obtenção de cada um deles. Outra forma de gerar a mesma economia em alguns casos é apelar para o senso comum. Por exemplo, no presente trabalho havia dois indicadores que obviamente estariam relacionados, o 17 e 18, sobre as redes de distribuição de água e acesso à água potável. Fosse o intuito desde o começo apenas realizar os agrupamentos e estudá-los, um destes dois indicadores poderia ser tranquilamente abandonado. O método ilustrado neste trabalho, entretanto, dá critérios para se tomar essa decisão com base na técnica e não apenas em senso comum.

Além disso, mostrou-se que há de se ter cuidado com o os indicadores a serem retirados e quais os critérios para fazê-lo. Mostrou-se que a retirada de indicadores com base no *biplot* PCA não é confiável para manutenção dos mesmos agrupamentos com menos indicadores sendo alimentados. Notou-se também que o patamar de 0,80 na matriz de correlação já pode ser considerado próximo o suficiente de 1, sendo poucos os casos, no exemplo estudado, que se encaixaram neste patamar (apenas 4,1% do total de correlações).

Por fim, recomendam-se alguns estudos adicionais: verificações das faixas de ângulo de vetores, variância total mantida e correlação necessárias a se manter os agrupamentos em conjuntos de dados mais diversos e consolidados, então partir para várias amostras de várias cidades e diversos indicadores para aumentar a confiabilidade dos resultados aqui demonstrados e também mostrar que as técnicas aqui escolhidas se comportam de forma similar quando substituídos os dados. A verificação de que, de fato, a análise de componente principal, técnica indicada no presente trabalho como a de mais valia para as análises de caso real, é a melhor para os mais diversos casos possíveis seria uma informação de muito valor a ser conseguida. Igualmente seriam o comportamento das técnicas bidimensionais como o dendrograma e o MDS em 2D. Não só o uso de indicadores, também se poderia experimentar a mesma metodologia sugerida neste trabalho para índices de inteligência de cidades. Além disso, a replicação deste mesmo estudo a um rol maior de cidades exclusivamente brasileiras de todos os portes, o que permitiria ao pesquisador brasileiro mais domínio do corpo de estudo, sendo sua experiência cotidiana refletida no trabalho. Outra linha possível de pesquisa que se abre é a de se tentar relacionar os dados quantitativos a dados qualitativos a partir de pesquisas de campo em escala menor. Dessa forma, seria possível trabalhar com pesquisas menores para suprir lacunas em dados ou mesmo partir inteiramente dessas para chegar-se a conclusões interessantes de agrupamento e diagnóstico de cidades inteligentes.

Notou-se ao longo de toda a labuta que gerou esta monografia que as cidades brasileiras têm muito para caminhar em termos de sustentabilidade, inteligência e resiliência.

7. REFERÊNCIAS BIBLIOGRÁFICAS

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

AHVENNIEMI, H. et al. What are the differences between sustainable and smart cities? **Cities**, Finlândia, 2017. 234-245.

ANDREWMAN327. Multidimensional scaling. **Wikipedia**, 2013. Disponível em: <<https://commons.wikimedia.org/wiki/File:RecentVotes.svg>>. Acesso em: 05 maio 2021.

ANTHOPOULOS, L. G. Understanding the Smart City Domain: A Literature Review. **Transforming City Governments for Successful Smart Cities**, 2015. p. 9-21.

BERNADINI, F. C. et al. General Features of Smart City Approaches from Information Systems Perspective and Its Challenges. **Sociedade Brasileira de Computação**, 2017. p. 25-40.

CHIUSOLI, C. L.; REZENDE, D. A. Indicadores para uma cidade inteligente e estratégica. **Revista Políticas Públicas & Cidades**, v. 8, n. 1, 2019.

CODEPLAN. BRASÍLIA/DF EM DADOS. **CODEPLAN**, 2019. Disponível em: <<http://brasiliadfemdados.codeplan.df.gov.br/>>. Acesso em: 7 out. 2021.

COUTO, E. D. A. **Aplicação dos indicadores de desenvolvimento sustentável da norma ABNT ISO 37120:2017 para a cidade do Rio de Janeiro e análise comparativa com cidades da América Latina**. UFRJ / Escola Politécnica. Rio de Janeiro, p. 163. 2018.

ETSI. **Key Performance Indicators for Sustainable Digital Multiservice Cities**. European Telecommunications Standards Institute. França, p. 62. 2017.

FAUSTO, B. **História do Brasil**. 14^a. ed. São Paulo: Editora da Universidade de São Paulo, 2019.

FLORES, L. E. B.; TEIXEIRA, C. S. Cidades Sustentáveis e Cidades Inteligentes: Uma análise dos rankings Arcadis e european smart cities. **Revista Eletrônica do Alto Vale do Itajaí**, v. 6, n. 9, p. 68-76, 2017.

G1. Economia. **G1**, 2021. Disponível em: <<https://g1.globo.com/economia/noticia/2021/04/23/sem-orcamento-censo-e-suspenso-mais-uma-vez-entenda-a-importancia-da-pesquisa-e-o-que-acontece-agora.ghtml>>. Acesso em: 7 out. 2021.

GIFFINGER, R. et al. **Ranking of European medium-sized cities**. Centre of Regional Science. Viena, p. 28. 2007.

GIFFINGER, R.; GUDRUN, H. Smart cities ranking: an effective instrument for the positioning of the cities? **ACE: architecture, city and environment**, v. 12, p. 7-25, fev. 2010.

GLEN, S. Multidimensional Scaling: Definition, Overview, Examples. **StatisticsHowTo**, 2015. Disponível em: <<https://www.statisticshowto.com/multidimensional-scaling/>>. Acesso em: 05 maio 2021.

GOULART, A. D. C. Revisitando a espanhola: a gripe pandêmica de 1918 no Rio de Janeiro. **História das Ciências da Saúde - Manguinhos**, Rio de Janeiro, abr. 2005. p. 101-142.

GUIMARÃES, R. P.; FEICHAS, S. A. Q. Desafios na construção de indicadores de sustentabilidade. **Ambiente & Sociedade**, Campinas, v. 12, n. 2, p. 307-323, dez. 2009.

HINTZE, J. L. **User's Guide IV: Multivariate Analysis, Clustering, MetaAnalysis, Forecasting/Time Series, Operations Research, Mass Appraisal**. EUA: NCSS, 2007.

HUOVILA, A.; BOSCH, P.; AIRAKSINEN, M. Comparative analysis of standardized indicators for Smart sustainable cities: What indicators and standards to use and when? **Cities**, v. 89, p. 141-153, 2019.

IBGE. Cidades. **IBGE**, 2017. Disponível em: <<https://cidades.ibge.gov.br/>>. Acesso em: 7 out. 2021.

ISO. **ISO 37120: Sustainable development of communities - Indicators for city services and quality of life**. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Suíça, p. 116. 2014.

ISO. **ISO 37122: Sustainable cities and communities – Indicators for smart cities**. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Suíça, p. 95. 2019.

KRAMA, M. R. Análise dos indicadores de desenvolvimento sustentável do Brasil, usando a ferramenta painel de sustentabilidade. **PUC/PR/ENGENHARIA DE PRODUÇÃO E SISTEMAS**, p. 185, 2009.

KRUSKAL, J. B. **Multidimensional scaling**. EUA: Sage, v. 11, 1978.

LOMBARDI, P. et al. Modelling the smart city performance. **Innovation: The European Journal of Social Science Research**, v. 25, n. 2, p. 137-149, 2012.

MAGALHÃES, M. T. Q. **Metodologia para Desenvolvimento de Sistemas de Indicadores: Uma Aplicação no Planejamento e Gestão da Política Nacional de Transportes**. Dissertação de Mestrado. Brasília: Departamento de Engenharia Civil e Ambiental, UnB. 2004. p. 135.

MALHEIROS, T. F.; PHILIPPI JR., A.; COUTINHO, S. M. V. Agenda 21 Nacional e Indicadores de Desenvolvimento Sustentável: contexto brasileiro. **Saúde e Sociedade**, São Paulo, v. 17, n. 1, p. 7-20, jan - mar 2008.

MARSAL-LLACUNA, M.-L.; COLOMER-LLINÀS, J.; MELÉNDEZ-FRIGOLA, J. Lessons in urban monitoring taken from sustainable and livable cities to better address the Smart Cities initiative. **Technological Forecasting & Social Change**, Girona, 2015. 611-622.

MATLAB. Dendrogram plot - MATLAB dendrogram. **Mathworks**. Disponível em: <<https://www.mathworks.com/help/stats/dendrogram.html>>. Acesso em: 18 abr. 2021.

NAM, T.; PARDO, T. A. **Conceptualizing smart city with dimensions of technology, people, and institutions**. Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in. Maryland: ACM Press. 2011. p. 282.

NEIROTTI, P. et al. Current trends in Smart City initiatives: Some stylised facts. **Cities**, v. 38, p. 25-36, 2014.

NGO, L. How to read PCA biplots and scree plots. **Bioturing**, 2018. Disponível em: <<https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/>>. Acesso em: 20 abr. 2021.

NIELSEN, F. Hierarchical Clustering. In: NIELSEN, F. **Introduction to HPC with MPI for Data Science**. Undergraduate Topics in Computer Science. ed. [S.l.]: Springer International Publishing, 2016. p. 195-211.

RIBEIRO, T. S. V. ISO 37120 e Objetivo do Desenvolvimento Sustentável 11: convergência frente à Agenda 2030. **Dissertação (Programa de Mestrado em Cidades Inteligentes e Sustentáveis) - Universidade Nove de Julho**, São Paulo, 2019. p. 84.

RIFFAT, S.; POWELL, R.; AYDIN, D. Future cities and environmental sustainability. **Future Cities and Environment**, 25 fev. 2016. p. 23.

RINGNÉR, M. What is principal component analysis? **Nature biotechnology**, v. 26, n. 3, p. 303-304, 2008.

SANTANA, E. D. S. D.; OLIVEIRA NUNES, É. D.; SANTOS, L. B. The use of ISO 37122 as standard for assessing the maturity level of a smart city. **International Journal of Advanced Engineering Research and Science**, v. 5, n. 12, p. 309-315, dez. 2018.

SANTANA, E. D. S.; NUNES, E. D. O. **Smart Cities Evaluations through SMM Framework - Sustainability Maturity Model**. 9th International Workshop on ADVANCES in ICT Infrastructures and Services. [S.l.]: [s.n.]. 2021.

SEGNESTAM, L. Environmental Performance Indicators. **Environment Department Papers**, Washington, n. 2, p. 52, Outubro 1999.

SICHE, R. et al. Índices versus indicadores: precisões conceituais na discussão da sustentabilidade de países. **Ambiente & sociedade**, v. X, n. 2, p. 137-148, jul-dez 2007.

SOBRAL, A. et al. Definições básicas : dado, indicador e índice. In: SAÚDE, B. -M. D. **Saúde ambiental: guia básico para construção de indicadores**. Brasília: Ministério da Saúde, 2017. p. 25-52.

THUSWOHL, M. Sociedade. **Carta Capital**, 2021. Disponível em: <<https://www.cartacapital.com.br/sociedade/favelas-do-rio-registram-mais-mortes-por-covid-19-que-muitos-paises/>>. Acesso em: 30 mar. 2021.

TOKARNIA, M. Um em cada 4 brasileiros não tem acesso à internet, mostra pesquisa. **Agência Brasil**, 2020. Disponível em: <<https://agenciabrasil.ebc.com.br/economia/noticia/2020-04/um-em-cada-quatro-brasileiros-nao-tem-acesso-internet/>>. Acesso em: 06 abr. 2021.

TORRES-SALINAS, D. et al. On the use of biplot analysis for multivariate bibliometric and scientific indicators. **Journal of the American Society for Information Science and Technology**, v. 64, n. 7, p. 1468-1479, 2013.

UNFPA. **State of World Population 2007 - Unleashing the Potential of Urban Growth**. UNITED NATIONS POPULATION FUND. Nova York, p. 108. 2007.

UNFPA. Urbanization. **UNITED NATIONS POPULATION FUND**, 2016. Disponível em: <<https://www.unfpa.org/urbanization>>. Acesso em: 20 mar. 2021.

WANG, J. **Geometric structure of high-dimensional data and dimensionality reduction**. Springer. ed. Beijing: Higher Education Press, 2012.