ANÁLISE FATORIAL GARANTIDA OU O SEU DINHEIRO DE VOLTA: UMA INTRODUÇÃO À REDUÇÃO DE DADOS[1] [2]

Dalson Britto Figueiredo Filho[3]
Enivaldo Carvalho da Rocha[4]
Ranulfo Paranhos[5]
Anderson Henrique Silva[6]
José Alexandre Silva Jr.[7]
Lucas Emanuel Oliveira[8]
Dáfni Priscila Alves[9]

Resumo
Esse artigo apresenta uma introdução intuitiva à analise fatorial. O objetivo principal é fornecer um guia prático para usuários com pouca familiaridade com o tema. Nosso público alvo são estudantes de graduação e pós-graduação em estágios iniciais de treinamento. Metodologicamente, o desenho de pesquisa utiliza simulação básica e replica duas bases de dados (PNUD, 2010 e Coopedge, Alvarez e Maldonado, 2008). Todas as análises estatísticas foram realizadas a partir do *Statistical Package for Social Sciences*. Com esse trabalho esperamos facilitar a compreensão e a aplicação de técnicas de redução de dados em Ciência Política.
Palavras-chave: redução de dados; análise fatorial; métodos quantitativos.

Abstract
This paper presents an intuitive introduction to factor analysis. The main purpose is to provide a applied guide to users with less knowledge on the subject. Our targeting audience are both undergraduate and graduate students in early state of training. Methodologically, the research design combines basic simulation and replicates two datasets (PNUD, 2010 and Coopedge, Alvarez and Maldonado, 2008). All statistical analyses were performed based on *Statistical Package for Social Sciences*. With this paper we hope to facilitate the understanding and the application of data reduction techniques in Political Science.
Keywords: data reduction; factor analysis; quantitative methods.

Resumen
Este artículo presenta una introducción intuitiva para el análisis factorial. El objetivo principal es proporcionar una guía práctica para los usuarios con poca familiaridad con el tema. Nuestro público objetivo es que los estudiantes de postgrado y estudios de postgrado en las primeras etapas de la formación. Metodológicamente, el diseño de la investigación utiliza réplicas de simulación básicos y

[1] DOI deste artigo: http://dx.doi.org/10.5380/recp.v5i2.40368
[2] Esse artigo foi elaborado a partir das notas de aula da disciplina Análises de Dados, ofertada pelo mestrado interinstitucional (MINTER) em Ciência Política através de um convênio UFPE/IFMT/CAPES. Agradecemos aos alunos pela participação ativa durante o curso e ao professor Michael Coopedge por compartilhar a sua base de dados. Eventuais limitações são monopólio dos autores. Material para replicação está disponível em: http://dx.doi.org/10.7910/DVN/27066
[3] Doutor em Ciência Política, Professor do Dept. de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE)
[4] Doutor em Engenharia de Produção, Professor do Dept. de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE)
[5] Doutor em Ciência Política, Professor do Instituto de Ciências Sociais de Universidade Federal de Alagoas (DCP/UFAL) Doutor em Ciência Política, Professor do Instituto de Ciências Sociais de Universidade Federal de Alagoas (DCP/UFAL)
[6] Graduando do Curso de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE)
[7] Doutor em Ciência Política, Professor do Instituto de Ciências Sociais de Universidade Federal de Alagoas (DCP/UFAL)
[8] Graduando do Curso de Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE)
[9] Mestranda do Programa de Pós-Graduação em Ciência Política da Universidade Federal de Pernambuco (DCP/UFPE)

dos bases de datos (PNUD, 2010 y Coopedge, Álvarez y Maldonado, 2008). Todos los análisis estadísticos se realizaron utilizando el paquete estadístico para las Ciencias Sociales. Con este trabajo esperamos facilitar la comprensión y la aplicación de técnicas de reducción de datos en Ciencias Políticas.
Palabras-clave: la reducción de datos; análisis factorial; Métodos cuantitativos.

There is no substitute for hard work
Thomas Edison

Sword of Omens, give me Sight Beyond Sight
Lion-O

1. INTRODUÇÃO

A análise fatorial é uma técnica de redução de dados. Mas o que isso quer dizer? Significa que ela deve ser utilizada para reduzir uma grande quantidade de variáveis observadas a um número menor de fatores/componentes. Essa redução se baseia no padrão de correlação observado entre as variáveis originais, e é explicada/representada pelos fatores/componentes. Por exemplo, suponha que o pesquisador deseja mensurar ideologia. Uma alternativa é elencar questões que teoricamente medem a variação da ideologia. Vejamos:

X_1. Você é a favor da descriminalização da maconha?
X_2. Você é a favor da união homoafetiva?
X_3. Você é a favor da adoção de crianças por casais homossexuais?
X_4. Você é a favor do aborto?
X_5. Você é a favor de programas de redistribuição de renda?

Assumindo uma codificação dicotômica (1 = Sim, 0 = Não), um indivíduo com escore 5 (é a favor de todas as assertivas) é considerado ideologicamente diferente de um respondente contrário a todas as alternativas (escore zero). Da forma como a codificação foi realizada, quanto menor o escore, maior o nível de conservadorismo. A análise fatorial pode ser utilizada para estimar um fator/componente de ideologia que não é diretamente observável, mas que causa/representa a variação das variáveis originais (X_1, X_2, X_3, X_4 e X_5). Desta forma, o objetivo da análise fatorial é gerar fatores/componentes não observados a partir das variáveis observadas (KING, 2001). Por exemplo, Lijphart (2003) utilizou análise fatorial para reduzir a dimensionalidade de dez variáveis observadas a duas dimensões latentes do desenho institucional: (1) Executivo-partidos e (2) Federal-unitária. Similarmente, Putnam (2005) utilizou análise fatorial para construir um índice de comunidade cívica e um indicador de desempenho institucional. É nesse sentido que essa técnica é especialmente útil para a construção de indicadores.

Esse artigo apresenta uma introdução intuitiva à analise fatorial. O principal objetivo é fornecer um guia prático para usuários com pouca familiaridade com o tema. Metodologicamente, o desenho de pesquisa utiliza simulação básica e replica duas diferentes bases de dados: (1) PNUD (2010) e (2) Coopedge, Alvarez e Maldonado (2008). Todas as análises estatísticas foram realizadas com auxílio do *Statistical Package for Social Sciences* (SPSS, versão 20) e as rotinas computacionais foram devidamente reportadas. Com esse trabalho esperamos facilitar a compreensão e aplicação das técnicas de redução de dados na pesquisa empírica em Ciência Política.

O restante do artigo está organizado da seguinte forma. A próxima seção apresenta os passos que devem ser seguidos para implementação de uma análise fatorial. Depois disso, apresentamos como as saídas computacionais devem ser interpretadas. A última seção sumariza as conclusões.

2. O PLANEJAMENTO DA ANÁLISE FATORIAL[10]

O planejamento de um desenho de pesquisa que utiliza análise fatorial deve seguir cinco estágios:

(1) definição da dimensão teórica de interesse, ou seja, da variável latente/construto;
(2) identificação das variáveis observadas que representam ou são causadas pela variável latente/construto/dimensão;
(3) coleta dos dados;
(4) implementação computacional e
(5) interpretação dos resultados.

1. Definição da dimensão teórica de interesse (variável latente/construto)

O Índice de Desenvolvimento Humano (IDH) procura representar, de forma sintética, o desenvolvimento humano utilizando não apenas a dimensão econômica, mas incorporando outros elementos que teoricamente estão associados a um melhor nível de

[10] Leitores interessados em aprofundar seus conhecimentos sobre redução de dados devem consultar as referencias bibliográficas. Em particular, para uma introdução não técnica ver Figueiredo Filho e Silva Júnior (2010) e Hair *et al.* (2009). Para textos introdutórios ver Kim e Mueller (1978a; 1978b) e Zeller e Carmines (1980). Para uma abordagem mais avançada ver Tabachinick e Fidell (2007). Para análise fatorial de dados *missing* ver Mackelprang (1970) e Ligny *et al.* (1981). Para análise fatorial de dados categóricos ver Bartholomew (1984) e Vermunt e Magidson (2005). Para aplicações práticas utilizando o *SPSS* ver Dancey e Reidy (2004), Pallant (2007) e Ho (2006).

bem estar, no caso, educação e saúde. Ou seja, como não é possível observar diretamente o nível de IDH de um determinado país uma alternativa é elencar variáveis que teoricamente representam a qualidade de vida, mas que podem ser diretamente observadas e mensuradas. Da mesma forma, ninguém observa diretamente o nível de democracia de um determinado sistema político. O que é diretamente observável são variáveis que teoricamente representam o conceito de democracia, por exemplo, existência de eleições livres e regulares. É nesse sentido que o primeiro passo para empregar a análise fatorial é escolher uma dimensão teórica de interesse. Depois disso, deve-se identificar as variáveis observadas que representam ou são causadas pelo construto/dimensão.

2. Identificação das variáveis observadas

Um dos principais desafios da pesquisa empírica é a operacionalização de variáveis, ou seja, a transformação de um determinado conceito em uma variável observada. Isso porque o mesmo conceito pode ser representado por diferentes variáveis. Por exemplo, imagine um desenho de pesquisa que procura mensurar a qualidade das políticas públicas. Naturalmente, a qualidade não pode ser diretamente observada. Dessa forma, o pesquisador deve identificar variáveis que teoricamente cumpram esse papel. Stein *et al.* (2008) elaboraram um dos modelos mais difundidos para a mensuração objetiva da qualidade das políticas em perspectiva comparada. O foco repousa sobre características do desenho institucional que teoricamente elevam a qualidade das políticas, independente do seu conteúdo. Vejamos: (1) estabilidade, (2) adaptabilidade; (3) coordenação e coerência; (4) qualidade de implementação; (5) orientação pública e (6) eficiência[11].

Nessa etapa o pesquisador enfrenta problemas de validade e confiabilidade de suas variáveis. A definição clássica de Nunnally (1967) postula que a confiabilidade diz respeito à consistência da mensuração, ou seja, o grau em que medidas repetidas sobre as mesmas unidades produzem resultados similares. Uma forma intuitiva de entender o conceito de confiabilidade é imaginar uma balança. Se a cada vez que o mesmo indivíduo subir na balança ela apontar valores diferentes, conclui-se que o instrumento não é

[11] A estabilidade é a capacidade de manter as políticas decididas ou de reforçar os acordos firmados. Adaptabilidade é a capacidade de tomar decisões e resolver problemas. Coordenação e coerência é a capacidade do governo de não "balcanizar" as políticas públicas ou de degenerar o governo em vários "subgovernos" com patrões e clientelas específicas. A qualidade da implementação é a capacidade de traduzir as políticas decididas em ações concretas. A orientação pública é a capacidade dos atores de implementar decisões que beneficiem a população de forma ampla e não apenas grupos específicos. Por último, a eficiência diz respeito a maximização do retorno dos gastos públicos em termos de resultados, ou seja, fazer mais com menos (Stein *et al.*, 2008).

confiável. Quanto maior a confiabilidade da medida, menor a quantidade de erro aleatório no processo de mensuração.

Por sua vez, a validade refere-se ao grau de correspondência entre o que se mediu e o que se queria medir (ZELLER e CARMINES, 19807; EVERITT e SKRONDAL, 2010). Nas palavras de Jannuzzi (2005), "validade é outro critério fundamental na escolha de indicadores, pois é desejável que se disponha de medidas tão próximas quanto possível do conceito abstrato ou da demanda política que lhes deram origem" (JANNUZZI, 2005: 139)[12]. Depois de identificar as variáveis e determinar como elas serão mensuradas o próximo passo é coletar os dados.

3. Coleta dos dados

Os dados podem ser primários, quando o próprio pesquisador produz a informação, ou secundários, quando o pesquisador trabalha informações coletadas por outros pesquisadores/instituições (IPEA, IBGE, etc.). O aumento da oferta *online* de informações representa uma oportunidade única para os cientistas políticos realizarem análises originais. Existem repositórios institucionais que ofertam gratuitamente bases de dados prontas para serem utilizadas. Por exemplo, no Brasil, o Consórcio de Informações Sociais (CIS) é um sistema de intercâmbio de informações científicas sobre a sociedade brasileira. Tem como objetivo oferecer gratuitamente dados qualitativos e quantitativos resultantes de pesquisas sobre vários aspectos da vida social[13]. No plano internacional tem-se o *Interuniversity Consortium for Political and Social Research* (ICPSR) que fornece acesso ao maior repositório de dados sociais do mundo[14]. Existe ainda o projeto Dataverse organizado pelo *The Institute for Quantitative and Social Science* da Universidade de Harvard[15].

Além disso, o avanço computacional facilitou os processos de coleta automatizada de dados. Existem diferentes formas de raspagem de dados (*web scraping*) que permitem economizar tempo e recursos, além de minimizar a probabilidade de erros. Por exemplo, *Silva et al.* (2015) desenvolveram um *software* especialmente voltado para coletar os dados do

[12] Para uma introdução à mensuração em Ciências Sociais ver Zeller e Carmines (1980). Para uma abordagem mais avançada ver Blalock (1979).
[13] O Consórcio de Informações Sociais é mantido pelo Departamento de Sociologia da Universidade de São Paulo (USP) e pela Associação Nacional de Pós-graduação e Pesquisa em Ciências Sociais (ANPOCS) e conta com suporte material e financeiro da USP e CNPq. Ver http://www.nadd.prp.usp.br/cis/
[14] Ver https://www.icpsr.umich.edu/icpsrweb/landing.jsp
[15] No original, "the Dataverse is an open source web application to share, preserve, cite, explore and analyze research data. It facilitates making data available to others, and allows you to replicate others work. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive appropriate credit". Ver http://dataverse.org/

Conselho Nacional de Justiça a respeito das sentenças condenatórias em casos de improbidade administrativa[16].

4. Implementação computacional

Depois de coletar os dados, deve-se escolher um *software* de análise de dados (SPSS, STATA, SAS, R, etc.) para implementar computacionalmente a análise fatorial. Primeiramente, deve-se observar o seguinte:

(1) o nível de mensuração das variáveis;

(2) tamanho da amostra e

(3) padrão de correlação (força) entre as variáveis observadas.

As variáveis devem ser preferencialmente contínuas e/ou discretas[17]. Em relação ao número de casos, deve-se evitar a utilização da análise fatorial com amostras pequenas já que os coeficientes de correlação estimados são menos confiáveis. Dessa forma, os fatores/componentes produzidos a partir de amostras pequenas tem menor poder de generalização (COSTELO e OSBORNE, 2005). Stevens (1996) argumenta que o número mínimo de casos recomendado pela literatura vem caindo ao longo do tempo. Nunnally (1978) sugere a utilização de 10 casos para cada variável incluída na análise. Logo, em uma base de dados com cinco variáveis, deve-se ter, no mínimo, 50 casos. Hair *et al.* (2009) sugerem 100 casos e/ou uma razão mínima de cinco casos por variável como uma situação aceitável. Tabachnick e Fidell (2007) sugerem 300 como tamanho mínimo. Costelo e Osborne (2005) fizeram uma revisão de 303 artigos e reportam que 62,9% dos trabalhos utilizam uma razão entre variáveis e casos de 10 ou menos. Nossa sugestão é de que quanto maior a amostra, melhor. Lembrando que a análise fatorial pode ser utilizada em amostras menores desde que o padrão de correlação ente as variáveis seja consistente. Ou seja, quanto maior o nível de correlação entre as variáveis incluídas, menor é a quantidade de casos para conseguir uma solução aceitável.

Por fim, no que diz respeito ao padrão de correlação entre as variáveis, a literatura recomenda que a maior parte dos coeficientes seja superior a 0,3, independente do sinal (TABACHNICK e FIDELL, 2007). Quanto maior o nível de correlação entre as variáveis observadas, mais adequadas são as diferentes técnicas de redução de dados. Existem outras

[16] O nome do aplicativo é TOGARY e os dados estão publicamente disponíveis no seguinte endereço eletrônico: http://dx.doi.org/10.7910/DVN/27787

[17] Variáveis ordinais e nominais também podem ser utilizadas, mas é preciso cautela na interpretação (HAIR *et al.*, 2009). Atualmente já existem modelos específicos para trabalhar com esse tipo de mensuração. Sugerimos ver a técnica de análise de correspondência e análise de classes latentes.

medidas que informam a adequabilidade dos dados. Por exemplo, a estatística de ajuste de Kaiser-Meyer-Olkin (KMO) indica a proporção da variância atribuída a um fator/componente comum[18]. Dessa forma, quanto mais perto de um, mais adequada é a aplicação da análise fatorial. O teste de esfericidade de Bartlett testa se variáveis são correlacionadas. Usualmente utiliza-se o nível de 5% de significância. Logo, quanto menor o p-valor do teste, maior a confiança em rejeitar a hipótese nula de que tem-se uma matriz identidade, ou seja, ausência de correlação.

Outro procedimento importante diz respeito à técnica de extração, ou seja, ao método utilizado para calcular os fatores/componentes. Matematicamente, a extração diz respeito à diagonalização da matriz. Isso porque a matriz de correlação entre as variáveis observadas pode ser diagonalizada de modo que os números na diagonal positiva representam a variância da matriz original. Existem diferentes formas de extrair os fatores/componentes[19] (componentes principais, mínimos quadrados não ponderados, mínimos quadrados generalizados, máxima verossimilhança, fatoração de eixo principal, fatoração alfa, fatoração imagem, entre outras). Aqui é importante analisar a diferença entre análise fatorial (AF) e análise de componentes principais (ACP). A ACP é mais popular e geralmente é utilizada como padrão em diferentes algoritmos computacionais. Ela utiliza toda a variância observada entre as variáveis e produz componentes que *representam* a variância das variáveis observadas. Por sua vez, a AF utiliza apenas a variância comum e produz fatores que causam as variáveis observadas Ambas procuram estimar combinações lineares das variáveis originais com o objetivo de explicar/representar a variância no padrão de correlação entre as variáveis.

Depois de optar pelo método de extração, o pesquisador deve definir o tipo de rotação e determinar quantos fatores/componentes devem ser extraídos. Mas para que serve a rotação? Resposta: simplificar e clarificar a estrutura dos dados (COSTELO e OSBORNE, 2005). Dito de outra forma, a rotação dos fatores/componentes procura facilitar a interpretação da solução observada sem alterar as suas propriedades matemáticas originais (TABACHNICK e FIDELL, 2007). Por exemplo, a rotação não pode melhorar a solução observada, ou seja, não adianta variar o método de rotação com o objetivo de

[18] Ver Tabachnick e Fidell (2007).
[19] Costelo e Osborne (2005) identificam as seguintes: *unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.*

aumentar a proporção de variância acumulada explicada pelo modelo. A literatura identifica duas principais formas de rotação: (1) ortogonais e (2) oblíquas[20].

A rotação ortogonal produz fatores/componentes não correlacionados. A matriz de autovalores (*loading matrix*) apresenta a correlação entre as variáveis observadas e os fatores/componentes extraídos. Na rotação ortogonal deve-se assumir que a correlação entre essas duas dimensões é zero, ou seja, elas são estatisticamente independentes (ortogonais). Por um lado, essa rotação é mais fácil de interpretar. Por outro, é mais difícil de assumir teoricamente a sua razoabilidade na medida em que pressupõe que os fatores/componentes extraídos são independentes. Por sua vez, a rotação oblíqua permite que os fatores/componentes sejam correlacionados. Dessa forma, ela é teoricamente mais razoável. A desvantagem é que a sua interpretação é mais complicada. A matriz de autovalores é particionada em duas: uma matriz de estrutura e uma matriz de padrão.

A determinação do número de fatores/componentes a serem extraídos é um dos principais problemas da análise fatorial (PREACHER *et al.*, 2013). Em geral, se tem duas diferentes abordagens: (1) empírica/exploratória e (2) teórica/confirmatória. A perspectiva exploratória é geralmente utilizada em estágios iniciais da pesquisa e/ou na ausência de teoria sobre o padrão de correlação esperado entre as variáveis. O pesquisador está interessado em testar hipóteses sobre o padrão de correlação observado entre as variáveis originais. Na perspectiva exploratória o autor é guiado por critérios empíricos de ajuste do modelo. Usualmente, os critérios mais recorrentes são a regra de Kaiser (*eigenvalue*), a análise gráfica do *Scree plot* e a análise da variância acumulada[21].

Por outro lado, a abordagem confirmatória usualmente é empregada em desenhos de pesquisa mais sofisticados que contam com algum pressuposto teórico mais consolidado. Aqui o pesquisador utiliza a teoria disponível para testar em que medida a solução encontrada por outros trabalhos se mantém consistente. Além disso, é possível estimar em que medida diferentes variáveis observadas podem ser reduzidas a um conjunto menor de fatores/componentes teoricamente inteligíveis.

[20] No caso da ortogonal ela pode ser do tipo *varimax*, *quartimax*, *equamax*, *parsimax*, entre outros. No caso da rotação oblíqua, as mais comumente utilizadas são *direct oblimin*, *direct quartimin, promax* e *procrustes*. Para uma análise detalhada de cada uma ver Tabachinick e Fidell (2007). Para outras conhecer outras formas de extração e outros métodos de rotação ver Mulaik (1972), Rummel (1970) e Gorsuch (1997).

[21] Pela regra de Kaiser, deve-se extrair apenas os fatores/componentes com autovalor maior do que um (K>1). Em relação ao *Scree plot*, deve-se observar a variação dos autovalores pelos fatores/componentes extraídos. Uma redução abrupta sugere que o pesquisador deve parar a extração e selecionar os componentes posicionados antes da quebra (COSTELO e OSBORNE, 2005).

5. Interpretação dos resultados

O último procedimento é interpretar substantivamente os resultados. O pesquisador deve discutir como os resultados se relacionam com o conhecimento acumulado em sua área específica de pesquisa. Em particular, quando o principal objetivo é estimar um indicador síntese, deve-se utilizar algum método para validar o componente/fator extraído. Por exemplo, um indicador válido de pobreza deve estar negativamente correlacionado com a renda per capita. Similarmente, um indicador válido de desempenho escolar deve estar positivamente correlacionado com as notas do Enem.

Por fim, importante é reportar todas as escolhas metodológicas adotadas com o objetivo de garantir a replicabilidade dos resultados. É impossível avaliar a validade das conclusões de um trabalho sem compreender integralmente como os resultados foram produzidos. Partindo desse pressuposto, a próxima seção apresenta as principais características do nosso desenho de pesquisa.

3. METODOLOGIA

Esta seção apresenta as principais características do desenho de pesquisa com o objetivo de garantir a replicabilidade dos resultados reportados (KING, 1995; BOYER, 2003; LUPIA e ELMAN, 2014; ISHIYAMA, 2014). O quadro 1 sumariza o tamanho da amostra, as variáveis, o tipo de extração e o número de fatores/componentes a serem extraídos.

QUADRO 1 - DESENHO DE PESQUISA

Exemplo	População	Variáveis[22]	Tipo Extração	Número fatores/componentes
Simulação	300 casos	x_1, x_2, x_3, x_4, e x_5.	ACP	?
PNUD (2010)[23]	5.565 municípios brasileiros	T_água, T_banágua, T_dens, T_lixo, T_luz, Água_esgoto, Parede	ACP	1
Coopedge, Alvarez e Maldonado (2008)	191 países	*Civil liberties, Political Rights, Index of Competition, Executive constraints, Comp. of Political participation, type of regime, Competitiveness of Executive Recruitment, Party Legitimacy, Legislative Effectiveness, Freedom of Assembly and Association, Freedom of Speech, Political Participation, Competitive Nomination Process, Adult Suffrage, Legislative Selection, Women's Political Rights, Effective Executive Selection, Index of Participation, and Openness of Executive Recruitment*	ACP e FA	2

Fonte: elaboração dos autores.

Com o objetivo de fazer valer o título do artigo, optamos por utilizar três diferentes exemplos sobre como a análise fatorial pode ser utilizada na pesquisa empírica. O primeiro se baseia em uma simulação que conta com uma amostra de 300 casos e examina o padrão de correlação entre cinco variáveis (x_1, x_2, x_3, x_4 e x_5). Utilizaremos a técnica de extração de análise de componentes principais sem definir *a priori* o número de componentes a serem extraídos.

O segundo exemplo replica dados do Programa das Nações Unidas para o Desenvolvimento (PNUD) sobre a condição de moradia nos 5.565 municípios brasileiros em 2010. As variáveis observadas são as seguintes: *T_água* representa percentual da população em domicílios com água encanada; *T_banágua* representa percentual da população em domicílios com banheiro e água encanada; *T_dens* representa o percentual da população que vive em domicílios com densidade superior a 2 pessoas por dormitório, *T_lixo* representa o percentual da população em domicílios com coleta de lixo, *T_luz* representa o percentual da população em domicílios com energia elétrica, *Água_esgoto* representa o percentual de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados e *Parede* representa o percentual de pessoas em

[22] Optamos por reportar exatamente os nomes das variáveis tal como aparecem nos bancos de dados originais para facilitar a reprodutibilidade das análises apresentadas.
[23] Os dados do PNUD (2010) estão disponíveis em http://www.atlasbrasil.org.br/2013/

domicílios com paredes inadequadas. Utilizaremos novamente a técnica de extração de análise de componentes e optamos por extrair um único indicador de qualidade de moradia (IQM)[24].

Por fim, o terceiro exemplo replica dos dados de Coopedge, Alvarez e Maldonado (2008) para estimar em que medida diferentes indicadores de democracia podem ser reduzidos às dimensões da poliarquia propostas por Dahl (1971): contestação e inclusividade. Originalmente, Coopedge, Alvarez e Maldonado (2008) utilizaram o método de análise de componentes principais e optaram pela rotação oblíqua que permite que os componentes extraídos sejam correlacionados. Manteremos o tipo de rotação original e ilustraremos a aplicação de um método alternativo de extração: fatoração do eixo principal (*principal factors*) com rotação ortogonal *varimax*. A próxima seção sumariza esses exemplos.

4. RESULTADOS

O primeiro passo é analisar a estatística descritiva e examinar matriz de correlação. A figura 1 sumariza essas informações.

FIGURA 1 - ESTATÍSTICA DESCRITIVA E MATRIZ DE CORRELAÇÃO[25]

Tabela 1 - Estatística descritiva

	média	**desvio padrão**	**n**
X_1	0	1	300
X_2	0	1	300
X_3	0	1	300
X_4	0	1	300
X_5	0	1	300

Tabela 2 - Matriz de correlação

	X_2	**X_3**	**X_4**	**X_5**
X_1	0,900	0,800	0,700	0,600
X_2	1	0,720	0,630	0,540
X_3		1	0,560	0,480
X_4			1	0,420
X_5				1

Fonte: elaboração dos autores.

Todas as variáveis foram simuladas com média zero e desvio padrão igual a um. Não existem dados ausentes (n = 300). Observa-se que o padrão de correlação entre as variáveis é consistente já que todos os coeficientes superam o patamar mínimo de 0,3 (ver tabela 2). Tem-se o primeiro indício de que a matriz de dados é adequada à utilização da análise fatorial A figura 2 apresenta as estatísticas de adequação da amostra e as comunalidades.

[24] Maiores detalhes sobre a criação do IQM podem ser encontrados em Figueiredo Filho *et al.* (2013; 2015).
[25] Analisar --> Redução de dimensão --> Fator. Em *Descritivos* deve-se selecionar as opções: descritivos de univariável, coeficientes, teste de esfericidade de Bartlett e KMO. Em *Extração* selecionar método de componentes principais e mostrar o *Scree plot*.

FIGURA 2 - MEDIDAS DE ADEQUAÇÃO DA AMOSTRA E COMUNALIDADES

Tabela 3 - Medidas de adequação da amostra

KMO	0,827
BTS (chi2)	1.127,30
(gl)	(10)
p-valor	0,000

Tabela 4 – Comunalidades

Variável	Extração
X_1	0,918
X_2	0,830
X_3	0,727
X_4	0,612
X_5	0,490

Fonte: elaboração dos autores.

O KMO foi de 0,827 (o máximo é 1). Já o teste de esfericidade foi estatisticamente significativo (p-valor<0,05). Logo, temos mais evidências para acreditar que a matriz é adequada (ver tabela 3). Por sua vez, as comunalidades representam a correlação entre o fator/componente e as variáveis originais (ver tabela 4). Quanto maior a correlação, maior é o nível de contribuição de uma determinada variável na criação do fator/componente. A literatura sugere que apenas variáveis com comunalidades acima de 0,4 devem permanecer na análise (SCHAWB, 2007). Para Costelo e Osborne (2005) uma comunalidade inferior a 0,4 sugere que talvez a variável não esteja correlacionada com as demais variáveis incluídas e/ou indica a existência de outro fator/componente. O próximo passo é examinar a variância total de cada fator/componente do modelo. A figura 3 apresenta essas informações.

FIGURA 3 - *SCREE PLOT* E VARIÂNCIA TOTAL EXPLICADA

Gráfico 1 - *Scree plot*

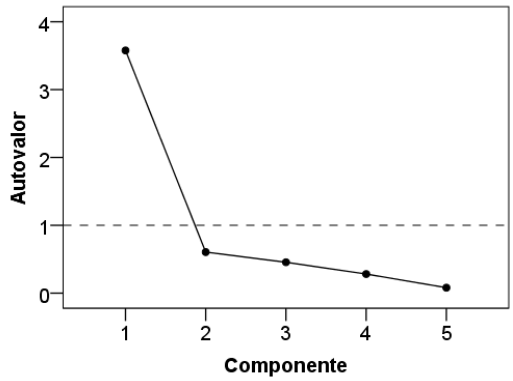


Tabela 5 - Variância total explicada

Componente	Total	% variância	% Acumulado
1	3,577	71,540	71,540
2	0,605	12,105	83,646
3	0,.456	9,110	92,755
4	0,.282	5,639	98,394
5	0,080	1,606	100,00

Fonte: elaboração dos autores.

O *Scree plot* (diagrama de sedimentação) ajuda a decidir quantos fatores devem ser extraídos (ver gráfico 1). A linha pontilhada representa o nível de corte da regra de Kaiser. Como pode ser observado, existe uma queda abrupta ao se comparar os componentes 1 e 2, ou seja, muita informação foi perdida. Seguindo a regra de Kaiser, apenas devemos

extrair os componentes com autovalor maior do que um. A solução observada sugere que o primeiro componente tem autovalor de 3,577 e carrega 71,540% da variância total das variáveis. Dito de outra forma, o primeiro componente representa 71,540% da informação total das variáveis observadas (ver tabela 5). Tanto a análise gráfica quanto a regra de Kaiser sugerem que apenas um componente deve ser extraído.

Em resumo, a solução encontrada apresenta as seguintes características: (1) as variáveis originais apresentam um padrão consistente de correlação; (2) tanto o KMO quanto o BTS indicam que a matriz é adequada para utilizar a análise fatorial; (3) todas as comunalidades superam o patamar mínimo de 0,4; e (5) o modelo explica 71,54% da variância observada. Esses resultados sugerem que a solução é tecnicamente correta e o fator/componente foi adequadamente extraído.

O segundo exemplo diz respeito às condições de moradia no Brasil. A tabela 6 apresenta a estatística descritiva das variáveis observadas.

TABELA 6 - ESTATÍSTICA DESCRITIVA (PNUD, 2010)

	média	**desvio padrão**	**N**
T_água	85,60	14,72	5.565
T_banágua	80,87	21,71	5.565
T_dens	25,13	13,00	5.565
T_lixo	94,05	11,05	5.565
T_luz	97,19	6,02	5.565
Água_esgoto	9,20	12,84	5.565
Parede	5,37	9,41	5.565

Fonte: elaboração dos autores.

Em 2010 a maior parte da população brasileira vivia em domicílios com oferta regular de água (média de 85,60% e desvio padrão de 14,72%). Mais de 1/4 da população vivia em domicílios com densidade superior a 2 pessoas por dormitório (*T_dens*). Os dados sobre coleta de lixo indicam que 94,05% da população contava com esse tipo de serviço. Por fim, quase 10% dos brasileiros viviam em domicílios com abastecimento de água e esgotamento sanitário inadequados. A tabela 7 ilustra o padrão de correlação entre as variáveis.

TABELA 7 - MATRIZ DE CORRELAÇÃO

	T_água	**T_banágua**	**T_dens**	**T_lixo**	**T_luz**	**Água_esgoto**	**Parede**
T_água	1						
T_banágua	0,719	1					
T_dens	-0,427	-0,750	1				
T_lixo	**0,278**	0,590	-0,473	1			
T_luz	0,408	0,582	-0,505	0,326	1		
Água_esgoto	-0,640	-0,874	0,684	-0,468	-0,423	1	
Parede	-0,248	-0,627	0,559	-0,611	-0,439	0,493	1

Fonte: elaboração dos autores

O padrão de correlação linear entre as variáveis é bastante consistente. Apenas um coeficiente ficou abaixou do patamar de 0,3, registre-se: a correlação entre *T_lixo* e *T_água* (r = 0,278). O próximo passo é examinar as estatísticas de ajuste da amostra e as comunalidades. A figura 4 sintetiza essas informações.

FIGURA 4 - ESTATÍSTICAS DE ADEQUAÇÃO DA AMOSTRA E COMUNALIDADES

Tabela 8 - Medidas de adequação da amostra

KMO	0,818
BTS (chi2)	27066,76
(gl)	(21)
p-valor	0,000

Tabela 9 – Comunalidades

Variável	Extração
T_água	0,473
T_banágua	0,918
T_dens	0,675
T_lixo	0,464
T_luz	0,442
Água_esgoto	0,744
Parede	0,530

Fonte: elaboração dos autores.

Novamente, tanto o KMO (0,818) quanto o BTS (p-valor<0,05) indicam que a matriz de dados é adequada à utilização da análise fatorial. Todas as variáveis apresentaram comunalidades acima do patamar mínimo sugerido pela literatura (COSTELO e OSBORNE, 2005). Comparativamente, *T_lixo* e *T_luz* são menos correlacionadas com o componente de qualidade de moradia. Uma possibilidade analítica é retirar variáveis com comunalidades reduzidas e estimar novamente o modelo em busca de melhores soluções. A figura 5 apresenta os componentes, autovalores e variância total carregada pelo modelo.

FIGURA 5 - *SCREE PLOT*, AUTOVALORES E VARIÂNCIA TOTAL EXPLICADA

Gráfico 2 - *Scree plot*

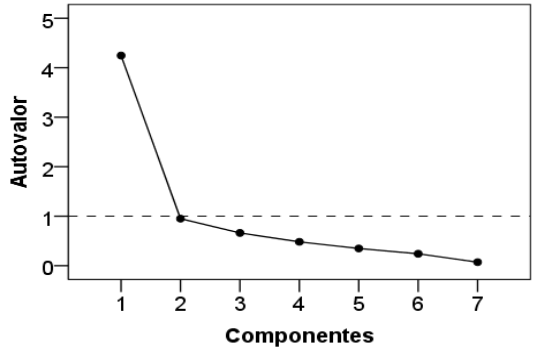


Tabela 10 - Variância total explicada

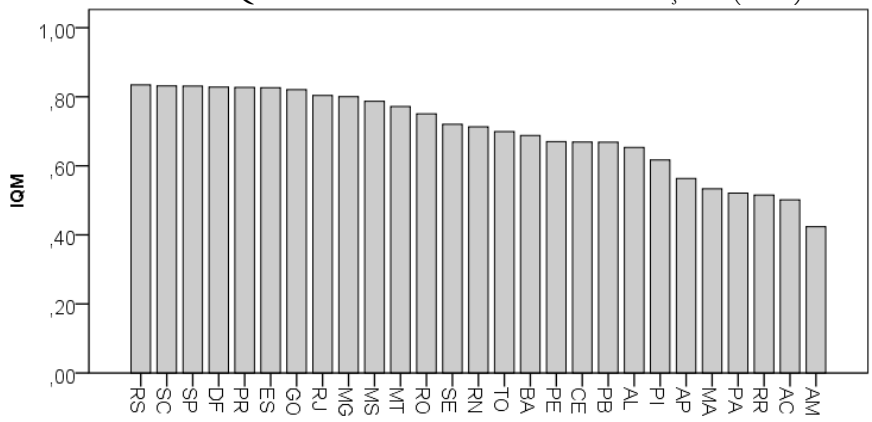
Componente	Total	% da variância	% Acumulado
1	4,25	60,66	60,66
2	0,95	13,57	74,23
3	0,66	9,46	83,69
4	0,48	6,88	90,57
5	0,35	4,97	95,55
6	0,24	3,43	98,98
7	0,07	1,02	100,00

Fonte: elaboração dos autores.

Em relação ao *Scree plot* observa-se uma queda abrupta ao se comparar os componentes 1 e 2, ou seja, muita informação foi perdida (ver gráfico 2). Seguindo a regra de Kaiser, apenas devemos extrair os componentes com autovalor maior do que um. A solução observada sugere que o primeiro componente tem um autovalor de 4,25 e carrega 60,66% da variância total das variáveis originais. Dito de outra forma, o primeiro componente representa 60,66% da informação total das variáveis observadas. Tanto a regra de Kaiser quanto a análise gráfica sugerem que apenas um componente deve ser extraído. Em termos substantivos, esse componente mensura a qualidade da moradia. Quanto maior, maior a qualidade das condições habitacionais.

Em resumo, a solução encontrada apresenta as seguintes características: (1) quase todas as variáveis apresentam um padrão consistente de correlação (exceção foi *T_lixo* x *T_água* com r = 0,278); (2) tanto o KMO quanto o BTS indicam que a matriz é adequada; (3) todas as comunalidades superaram o patamar mínimo de 0,4; e (5) o componente extraído representa 60,66% da variância total. Depois de extraído, o componente/fator pode ser utilizado como variável dependente ou independente em modelos de regressão, por exemplo. Aqui, optamos por analisar descritivamente como o Índice de Qualidade de Moradia varia por unidade da federação. O gráfico 3 ilustra essas informações.

GRÁFICO 3 - IQM POR UNIDADE DA FEDERAÇÃO (2010)[26]



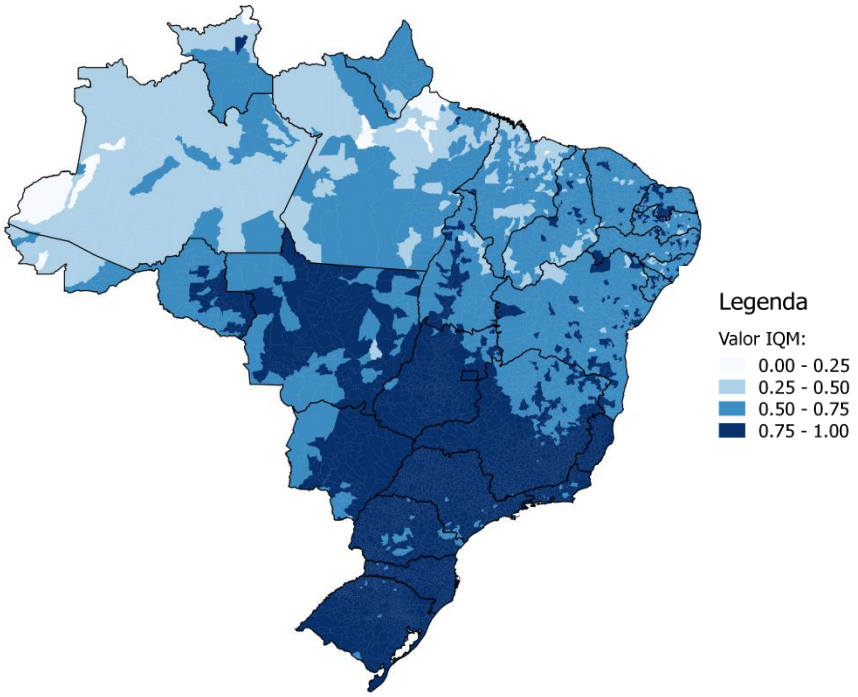
Fonte: elaboração dos autores

Para facilitar a interpretação, o indicador foi normalizado entre 0 e 1. Quanto mais perto de um, maior é a qualidade da moradia. Comparativamente, enquanto Rio Grande do

[26] Para salvar os componentes/fatores na matriz de dados do SPSS deve-se selecionar Pontuações --> salvar como método de regressão. Originalmente, os componentes/fatores extraídos são padronizados, ou seja, tem média zero e desvio padrão igual a um.

Sul (0,835), São Paulo (0,831) e Santa Catarina (0,831) apresentam as melhores condições de moradia, Roraima (0,515), Acre (0,501) e Amazonas (0,424) demonstram a maior vulnerabilidade habitacional. O Mapa 1 ilustra a variação do IQM por município.

MAPA 1 - IQM POR MUNICÍPIO (2010)



Fonte: elaboração dos autores.

A distribuição foi dividida em quatro grupos de mesmo tamanho. Quanto mais forte a cor, maior é o nível da qualidade da moradia. Comparativamente, observa-se que as regiões Sul e Sudeste apresentam melhores condições habitacionais do que Norte e Nordeste.

Por fim, o terceiro exemplo replica dados de Coopedge, Alvarez e Maldonado (2008). A tabela 11 apresenta a estatística descritiva das variáveis.

TABELA 11 - ESTATÍSTICA DESCRITIVA

Variável	média	desvio padrão	N
Civil Liberties	4,43	2,04	125
Political Rights	4,29	2,26	125
Competition	22,02	24,87	125
Executive Constraints (Decision Rules)	3,58	2,35	125
Political participation	0,93	0,90	125
Type of Regime	0,66	0,47	125
Competitiveness of Executive Recruitment	1,56	1,11	125
Party Legitimacy	3,89	4,31	125
Legislative Effectiveness	5,25	3,15	125
Freedom of assembly and association	0,82	0,88	125
Freedom of speech	0,90	0,79	125
The Competitiveness of Participation	2,46	1,62	125
Adult Suffrage (%)	78,50	39,08	125
Legislative Selection	8,84	3,05	125
Women's political rights	1,61	0,67	125
Effective Executive Selection	5,04	3,89	125
Participation	29,59	23,72	125
Openness of Executive Recruitment	3,18	1,53	125

Fonte: elaboração dos autores.

Em relação ao número de casos, apesar da amostra original conter 191 países, a análise foi realizada a partir de 125 casos já que o SPSS exclui os casos ausentes como procedimento padrão[27]. A tabela 12 ilustra os testes de adequação da amostra.

TABELA 12 - MEDIDAS DE ADEQUAÇÃO DA AMOSTRA

KMO	0,932
BTS (chi2)	2.885,10
(gl)	(153)
p-valor	0,000

Fonte: elaboração dos autores.

O KMO foi 0,932 (bastante próximo de 1) e o BTS foi significativo (p-valor<0,05). Isso indica que devemos rejeitar a hipótese nula de que estamos trabalhando com uma matriz identidade, ou seja, que as variáveis não são correlacionadas. A tabela 13 sintetiza as comunalidades.

TABELA 13 - COMUNALIDADES

Variável	Extração
Civil Liberties	0,910
Political Rights	0,929
Competition	0,876
Executive Constraints (Decision Rules)	0,874
Political participation	0,817
Type of Regime	0,801

[27] Uma opção é substituir os valores ausentes pela média de cada variável. Outra saída é trabalhar com alguma técnica específica de imputação de dados. Para os interessados no assunto ver King *et al.* (2001) e Honaker e King (2010).

Competitiveness of Executive Recruitment	0,869
Party Legitimacy	0,888
Legislative Effectiveness	0,867
Freedom of assembly and association	0,783
Freedom of speech	0,684
The Competitiveness of Participation	0,928
Adult Suffrage (%)	0,800
Legislative Selection	0,673
Women's political rights	**0,466**
Effective Executive Selection	**0,436**
Participation	0,580
Openness of Executive Recruitment	**0,450**

Fonte: elaboração dos autores.

Todas as variáveis apresentaram comunalidades acima do patamar mínimo de 0,4. Observa-se que *Women's political rights* (0,466), *Effective Executive Selection* (0,436) e *Openness of Executive Recruitment* (0,450) tem comunalidades mais reduzidas. Comparativamente, isso quer dizer que essas variáveis contribuem menos para a criação do fator/componente. A tabela 14 sumariza a variância total explicada.

TABELA 14 - VARIÂNCIA TOTAL EXPLICADA

Componente	Total	% variância	% Acumulado
1	**11,45**	**63,62**	**63,62**
2	**2,18**	**12,11**	**75,73**
3	0,83	4,60	80,32
4	0,72	3,98	84,30
5	0,58	3,20	87,50
6	0,50	2,80	90,30
7	0,40	2,20	92,50
8	0,25	1,37	93,87
9	0,22	1,23	95,10
10	0,21	1,19	96,29
11	0,15	0,81	97,10
12	0,13	0,70	97,80
13	0,10	0,55	98,35
14	0,09	0,49	98,84
15	0,07	0,39	99,23
16	0,06	0,34	99,57
17	0,04	0,25	99,82
18	0,03	0,18	100,00

Fonte: elaboração dos autores.

O critério de Kaiser indica que devemos extrair dois componentes. O primeiro tem autovalor de 11,45 e representa 63,62% da variância das variáveis originais. O segundo componente tem autovalor de 2,18 e representa 12,11%. Em conjunto, os dois componentes carregam 75,73% da variância total. O próximo passo é observar como os componentes extraídos e as variáveis observadas se relacionam.

TABELA 15 - MATRIZ DE COMPONENTES

Variável	1	2
Civil Liberties	**-0,943**	0,200
Political Rights	**0,943**	-0,198
Competition	**0,933**	-0,136
Executive Constraints (Decision Rules)	**0,930**	
Political participation	**0,928**	
Type of Regime	**0,924**	-0,143
Competitiveness of Executive Recruitment	**0,922**	-0,165
Party Legitimacy	**-0,915**	0,269
Legislative Effectiveness	**0,902**	
Freedom of assembly and association	**-0,870**	0,209
Freedom of speech	**0,866**	-0,181
The Competitiveness of Participation	**0,804**	-0,194
Adult Suffrage (%)	**0,635**	**0,419**
Legislative Selection	**0,554**	**0,379**
Women's political rights	0,565	**0,693**
Effective Executive Selection	0,475	**0,669**
Participation	0,347	**0,588**
Openness of Executive Recruitment	0,442	**0,490**

Fonte: elaboração dos autores.

Na matriz de componentes, deve-se observar a relação entre variáveis observadas e fatores/componentes extraídos. Em uma pesquisa substantiva, seria possível questionar porque a maior parte das variáveis se concentrou no componente 1. Ou, seria possível inquirir porque a apresentação original dos dados suprimiu as cargas fatoriais de ambos os componentes. No entanto, como nosso artigo é pedagógico, o mais importante aqui é analisar tecnicamente a solução observada. O pressuposto da estrutura simples determina que nenhuma variável pode ter carga fatorial alta em ambos os fatores/componentes. Isso porque fica impossível saber se a variável contribui para um ou para o outro. Por exemplo, a variável *Adult Suffrage (%)* que teoricamente mede a dimensão da inclusão apresentou carga fatorial alta em ambos os componentes, registre-se (0,635) no componente 1 e 0,419 no componente 2. Dessa forma, quanto mais variáveis apresentarem cargas fatoriais acima de 0,4 em ambos os componentes, mais o pressuposto da estrutura simples foi violado.

Observa-se que as variáveis *Civil Liberties, Political Rights, Competition, Executive Constraints (Decision Rules), Political participation, Type of Regime, Competitiveness of Executive Recruitment, Party Legitimacy, Legislative Effectiveness, Freedom of assembly and association, Freedom of speech, The Competitiveness of Participation, Adult Suffrage (%) e Legislative Selection* se relacionam mais fortemente com o primeiro componente. Em termos substantivos, essa dimensão representa o que Dahl (1971) chamou de contestação. Por sua vez, as variáveis *Women's political rights, Effective Executive Selection, Participation e Openness of Executive Recruitment* apresentam carga fatorial mais elevada no segundo componente, o que Dahl (1971) denominou de inclusividade.

Depois de replicar a solução proposta por Coopedge, Alvarez e Maldonado (2008), optamos por estimar um novo modelo com as seguintes características: (1) método de extração de fatoração por eixo principal; (2) rotação ortogonal varimax; e (3) casos ausentes substituídos pela média. A tabela 16 apresenta a estatística descritiva.

TABELA 16 - ESTATÍSTICA DESCRITIVA

Variável	média	desvio padrão	N	ausentes
Civil Liberties	4,32	2,00	170	5
Political Rights	4,16	2,22	170	5
Competition	22,68	24,25	170	8
Executive Constraints (Decision Rules)	3,54	2,11	170	34
Political participation	0,91	0,80	170	34
Type of Regime	1,54	0,99	170	34
Competitiveness of Executive Recruitment	0,64	0,47	170	7
Party Legitimacy	4,13	4,43	170	3
Legislative Effectiveness	5,62	3,23	170	1
Freedom of assembly and association	0,83	0,80	170	31
Freedom of speech	0,93	0,70	170	34
The Competitiveness of Participation	2,38	1,45	170	34
Adult Suffrage (%)	76,12	40,22	170	2
Legislative Selection	8,91	2,90	170	1
Women's political rights	1,60	0,60	170	34
Effective Executive Selection	4,85	3,63	170	1
Participation	29,24	22,45	170	8
Openness of Executive Recruitment	3,15	1,38	170	34

Fonte: elaboração dos autores.

Em relação ao número de casos, apesar da amostra original conter 191 países, a análise foi realizada a partir de 170 casos[28]. A tabela 17 ilustra os testes de adequação da amostra.

TABELA 17 - MEDIDAS DE ADEQUAÇÃO DA AMOSTRA

KMO	0,928
BTS (chi2) (gl)	2.481,29
p-valor	0,000

Fonte: elaboração dos autores.

O KMO foi de 0,928 (bastante próximo de 1) e o BTS foi significativo (p-valor<0,05). Isso indica que devemos rejeitar a hipótese nula de que estamos trabalhando com uma matriz identidade, ou seja, que as variáveis analisadas não são correlacionadas. A tabela 18 sumariza as comunalidades.

[28] Uma das vantagens da imputação é reduzir a quantidade de casos ausentes. Uma das desvantagens é reduzir a variabilidade da variável original.

TABELA 18 - COMUNALIDADES

Variável	Extração
Civil Liberties	0,860
Political Rights	0,893
Competition	0,848
Executive Constraints (Decision Rules)	0,789
Political participation	0,731
Type of Regime	0,786
Competitiveness of Executive Recruitment	0,737
Party Legitimacy	0,760
Legislative Effectiveness	0,750
Freedom of assembly and association	0,690
Freedom of speech	0,577
The Competitiveness of Participation	0,814
Adult Suffrage (%)	0,725
Legislative Selection	0,544
Women's political rights	**0,264**
Effective Executive Selection	**0,369**
Participation	0,535
Openness of Executive Recruitment	**0,306**

Fonte: elaboração dos autores.

A maior parte das variáveis apresentou comunalidades acima do patamar mínimo de 0,4. *Women's political rights* (0,264), *Effective Executive Selection* (0,369) e *Openness of Executive Recruitment* (0,306) apresentam comunalidades mais reduzidas, repetindo o padrão anteriormente observado quando empregamos análise de componentes principais como técnica de extração. Em uma análise substantiva, essas variáveis deveriam ser excluídas da análise e o pesquisador deveria estimar um novo modelo. A tabela 19 sumariza a variância total explicada.

TABELA 19 - VARIÂNCIA TOTAL EXPLICADA

Componente	Total	% variância	% Acumulado
1	**10,54**	**58,57**	**58,57**
2	**2,14**	**11,86**	**70,43**
3	0,94	5,20	75,63
4	0,80	4,47	80,10
5	0,75	4,16	84,26
6	0,51	2,86	87,12
7	0,45	2,48	89,59
8	0,34	1,88	91,48
9	0,30	1,67	93,15
10	0,24	1,35	94,49
11	0,23	1,29	95,79
12	0,19	1,06	96,85
13	0,16	,88	97,72
14	0,13	,70	98,42
15	0,11	,59	99,01

16	0,07	,41	99,42
17	0,07	,36	99,78
18	0,04	,22	100,00

Fonte: elaboração dos autores.

O critério de Kaiser indica que devemos extrair dois fatores. O primeiro tem autovalor de 10,54 e explica 58,57% da variância das variáveis originais. O segundo fator tem autovalor de 2,14 e explica 11,86%. Em conjunto, os dois fatores explicam 70,43% da variância total, valor muito próximo ao observado anteriormente na solução utilizando análise de componentes principais que carregou 75,73% da variância total. O próximo passo é observar como os fatores extraídos e as variáveis observadas se relacionam.

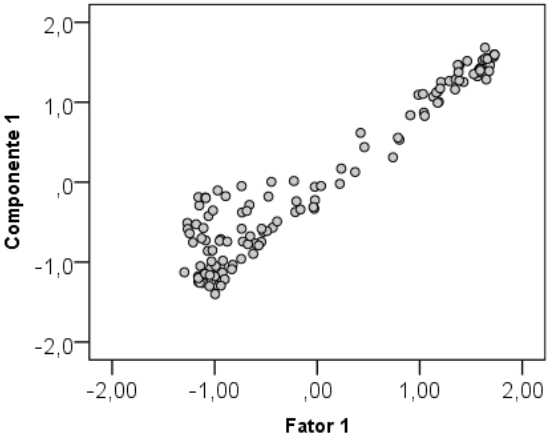
TABELA 20 - MATRIZ DE COMPONENTES

Variável	**1**	**2**
Civil Liberties	-0,890	0,262
Political Rights	-0,927	0,184
Competition	0,911	-0,132
Executive Constraints (Decision Rules)	0,875	-0,153
Political participation	0,855	0,011
Type of Regime	0,879	-0,112
Competitiveness of Executive Recruitment	-0,845	0,151
Party Legitimacy	0,866	-0,102
Legislative Effectiveness	0,865	0,049
Freedom of assembly and association	0,819	-0,138
Freedom of speech	0,749	-0,129
The Competitiveness of Participation	0,884	-0,181
Adult Suffrage (%)	0,507	0,684
Legislative Selection	0,450	0,584
Women's political rights	0,311	0,409
Effective Executive Selection	0,386	0,469
Participation	0,592	0,429
Openness of Executive Recruitment	0,501	0,235

Fonte: elaboração dos autores.

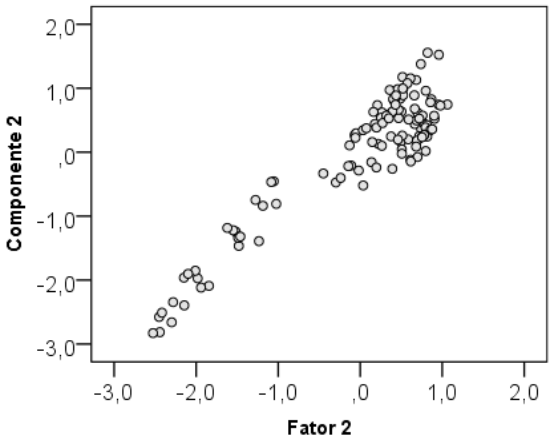
Para garantir a transparência dos resultados, optamos por reportar o nível de correlação entre os componentes e fatores extraídos nas duas soluções anteriormente apresentadas. Os gráficos 4 e 5 ilustram essas informações.

GRÁFICO 4 - SOLUÇÃO ORIGINAL (PCA)



r = 0,958
p-valor<0,001
n = 125

GRÁFICO 5 - SOLUÇÃO ALTERNATIVA (AF)



r = 0,925
p-valor<0,001
n = 170

Fonte: elaboração dos autores.

A solução original utilizou o método de análise de componentes principais, rotação oblíqua e exclusão de casos ausentes. A solução alternativa empregou fatoração do eixo principal, rotação ortogonal varimax e substituiu os casos ausentes pela média. Como pode ser observado, a correlação entre o fator 1 e o componente 1 é de 0,958 (p-valor<0,001), já a correlação entre o fator 2 e o componente 2 é de 0,925 (p-valor<0,001). Em termos menos técnicos, isso quer dizer que soluções distintas produziram resultados bastante semelhantes.

5. CONSIDERAÇÕES FINAIS

Esse artigo enfrentou um desafio: ofertar uma introdução intuitiva à redução de dados. Nosso principal objetivo foi fornecer um guia prático para usuários com pouca familiaridade com o tema. Nosso público alvo são estudantes de graduação e pós-graduação em fases iniciais de treinamento. Metodologicamente, o desenho de pesquisa utilizou simulação básica e replicou as bases de dados do PNUD (2010) e de Coopedge, Alvarez e Maldonado (2008). As análises estatísticas foram realizadas com auxílio do *Statistical Package for Social Sciences* (SPSS, versão 20) e as rotinas computacionais foram devidamente reportas. Dessa forma, usuários iniciantes poderão seguir o passo a passo computacional e implementar as suas próprias análises. Com esse trabalho esperamos

facilitar a compreensão e a aplicação de técnicas de redução de dados na pesquisa empírica em Ciência Política.

REFERÊNCIAS BIBLIOGRÁFICAS

BARTHOLOMEW, D. J. (1984). The foundations of factor analysis, *Biometrika*, 71, 221-232.

BLALOCK, H. M. (1979). The Presidential Address: Measurement and Conceptualization Problems: The Major Obstacle to Integrating Theory and Research. *American Sociological Review,* Vol. 44, No. 6 (Dec., 1979), pp. 881-894.

BOYER, M. A. (2003). Symposium on replication in International studies research. *International Studies Perspectives* 4, 72-107.

COPPEDGE, M.; ALVAREZ, A.; MALDONADO, C. Dimensions of Democracy: Contestation and Inclusiveness". *Journal of Politics*, v. 70, n. 3, p. 145, 2008.

COSTELLO, A. B. and OSBORNE, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10, 7, 13-24.

DAHL, R. (1971). *Poliarchy: Participation and Opposition,* New Haven, Yale University Press.

DANCEY, C.; REIDY, J. (2006). *Estatística Sem Matemática para Psicologia: Usando SPSS para Windows*. Porto Alegre: Artmed.

EVERITT, B. S.; SKRONDAL, A. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press, 449p.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. da (2010). Visão além do alcance: uma introdução à análise fatorial. *Opinião Pública*, v. 16, n. 1, p. 160-185.

FIGUEIREDO FILHO, D. B.; PARANHOS, R.; ROCHA, E. C. da; SILVA Jr., J. A da; MAIA, R. Análise de componentes principais para construção de indicadores sociais. *Rev. Bras. Biom.*, São Paulo, v.31, n.1, p.61-78, 2013

GORSUCH, R. L. Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, 68(3), 532-560, 1997.

HAIR, JR *et al.* (2009). *Multivariate Data Analysis*. 6. ed. Upper Saddle River: Pearson Prentice Hall.

HONAKER, J.; KING, G. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, vol. 54, n. 2, p. 561-581, 2010.

ISHIYAMA, J. (2014). "Replication, Research Transparency, and Journal Publications: Individualism, Community Models, and the Future of Replication Studies." *PS: Political Science & Politics*, 47 (01): 78–83.

JANNUZZI, P. de M. (2005). Indicadores para diagnóstico, monitoramento e avaliação de programas sociais no Brasil. *Revista do Serviço Público.* Brasília 56 (2). Abr/Jun, p. 137-160.

KIM, J. and MUELLER, C. W. (1978a). *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage.

KIM, J. and MUELLER, C. W. (1978b). *Introduction to factor analysis - what it is and how to do it.* Beverly Hills, CA: Sage.

KING, G. (1995). "Replication, Replication." *PS: Political Science and Politics,* 28: 443-499. Disponível: http://gking.harvard.edu/gking/files/replication.pdf

KING, G. (2001). How not to lie with statistics [Online]. Disponível em: <http://gking.harvard.edu/files/mist.pdf> Acesso em: 18 fev. 2015.

KING, G.; HONAKER, J.; JOSEPH, A.; and SCHEVE, K. 2001. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95(1): 49–69. Disponível em: <http://gking .harvard.edu/files/abs/evil-abs.shtml> Acesso em: 12 fev. 2015.

LIGNY, C. L.; NIEUWDORP, G. H. E; BREDERODE, W. K; HAMMERS, W. E; and HOUWELINGEN, J. C. van. (1981). An Application of Factor Analysis with Missing Data. *Technometrics*, 23, 1, 91-95.

LIJPHART, A. *Modelos de Democracia: desempenho e padrões de governo em 36 países.* Rio de Janeiro: Civilização Brasileira, 2003.

LUPIA, A., and ELMAN, C. (2014). Openness in Political Science: Data Access and Research Transparency. *PS: Political Science & Politics*, 47 (01): 19–42.

MACKELPRANG, A. J. (1970). Missing Data in Factor Analysis and Multiple Regression. *Midwest Journal of Political Science*, 14, 3, 493-505.

MULAIK, S. A. (1972). The foundations of factor analysis. New York, McGraw-Hill.

NUNNALLY, J. (1967). *Psychometric Methods.* New York: MacGraw Hill.

PALLANT, J. (2007). *SPSS Survival Manual*. Open University Press.

PREACHER, K. J., RUCKER, D. D., & HAYES, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42, 185–227.

PUTNAM, R. (2005). Comunidade e democracia: a experiência da Itália moderna. 4. Ed. - Rio de Janeiro: FGV.

RUMMEL, R. J. (1970). *Applied Factor Analysis*. Evanston: Northwestern University Press.

SCHAWB, A. J. (2007). Eletronic Classroom. [Online] Disponível em: <http://www.utexas.edu/ssw/eclassroom/schwab.html> Acesso em: [22 jan. 2010].

STEIN, E. et al. (2008). *Policymaking in Latin America. How Politics Shapes Policies*. Cambridge, United States: Harvard University, David Rockefeller Center for Latin American Studies.

STEVENS, J. (1996). *Applied multivariate for the social sciences*. 3. ed. Mahwah, NJ: Lawrence Erlbaum.

TABACHNICK, B. and FIDELL, L. (2007). *Using multivariate analysis*. Needham Heights: Allyn & Bacon.

VERMUNT, J. K.; MAGIDSON, J. (2004). "Factor Analysis with categorical indicators: A comparison between traditional and latent class approaches". *In*: VAN DER ARK, A.; CROON, M.A. e SIJTSMA, K. *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Mahwah: Erlbaum.

ZELLER, R. A. & CARMINES, E. G. (1980). *Measurement in the social sciences: The link between theory and data*. Cambridge: Cambridge University Press.

ANEXOS

QUADRO 2 - SIMULAÇÃO PASSO A PASSO

Procedimento
Criação de duas variáveis (A e B) com média zero e desvio padrão igual a um COMPUTE a=RV.NORMAL(0,1). EXECUTE. COMPUTE b=RV.NORMAL(0,1). EXECUTE.
A partir de A e B, realizamos uma análise de componentes principais e salvamos dois componentes (x e y). A correlação entre os componente é zero (r = 0). FACTOR /VARIABLES a b /MISSING LISTWISE /ANALYSIS a b /PRINT INITIAL EXTRACTION /CRITERIA FACTORS(2) ITERATE(25) /EXTRACTION PC /ROTATION NOROTATE /SAVE REG(ALL) /METHOD=CORRELATION.
Definimos x_1, x_2, x_3, x_4 e x_5 a partir da combinação linear de x e y, de modo a satisfazer as seguintes correlações: x1 x2 = 0,9; x1 x3 = 0,8; x1 x4 = 0,7 e x1 x5 = 0,6. COMPUTE x1=x * 1 + y * SQRT(1-1 ** 2). EXECUTE. COMPUTE x2=x1 * .9 + y * SQRT(1-.9 ** 2). EXECUTE. COMPUTE x3=x1 * .8 + y * SQRT(1-.8 ** 2). EXECUTE. COMPUTE x4=x1 * .7 + y * SQRT(1-.7 ** 2). EXECUTE. COMPUTE x5=x1 * .6 + y * SQRT(1-.6 ** 2).