Seleção dos indicadores mais relevantes para melhorar o monitoramento dos dados em estatal de transporte de passageiros utilizando PCA

Daniel Alexandre da Silva Gomes (USP) daniel.alexandre.eng@gmail.com
Miguel Ângelo Lellis Moreira (UFF) miguellellis@hotmail.com
Jarbas Honorio de Miranda (USP) jhmirand@usp.br
Luiz Paulo Fávero (USP) lpfavero@usp.br
Marcos dos Santos (IME) marcosdossantos@ime.eb.br

Resumo

O Monitoramento efetivo de indicadores realizado pelas empresas é crucial para uma gestão otimizada, para alocação de recursos devidamente e promoção da melhoria do desempenho. No caso da empresa estudada, o resultado obtido foi essencial, principalmente por se tratar da utilização de verba pública. Para isto, este trabalho teve o propósito de aplicar a técnica de aprendizado de máquina chamada PCA para a seleção de indicadores de desempenho efetivos para a gestão, provenientes de um conjunto de indicadores pré-estabelecidos pela empresa. Os dados analisados foram fornecidos pelo setor operacional a fim de chegar a uma quantidade de reduzida de indicadores, mas que fosse capaz de traduzir para os gestores informações suficientes sem a necessidade de esforço para analisar os dados original. Este estudo trouxe uma significativa contribuição para a empresa porque os resultados obtidos permitiram avaliar os indicadores de custo, segurança e manutenção para uma importante tomada de decisão baseada em dados, melhorando o desempenho corporativo.

Palavras-Chaves: KPI; indicadores; visualização de dados; tomada de decisão; redução de dimensionalidade; performance.

1. Introdução

A maioria da população mundial é urbana e essa parcela da população tende a continuar aumentando (UN, 2019). Este fenômeno acarreta carência de acesso a serviços básicos além de empregos e desenvolvimento profissional, os quais são afetados pela mobilidade urbana, sobretudo por transporte público de massa em contraponto aos veículos motorizados individuais (HOBBS et al., 2021). No Brasil, 83% da população vive em áreas urbanas, com

66% da população economicamente ativa dependendo do transporte coletivo e 22% desta parcela gastando mais de duas horas diárias no trânsito, refletindo na economia do País (CLARK, 2018). Priorizar o transporte público coletivo pode trazer benefícios econômicos e sociais (congestionamentos, acidentes e exclusão social) bem como ambientais (HOBBS *et al*., 2021). De acordo com ANPTrihos (2018) as vantagens do transporte sobre trilhos são a sua alta capacidade, previsibilidade, diminuição dos congestionamentos, favorecimento do desenvolvimento econômico e democratiza o acesso a outros direitos constitucionais.

Como empresa do Governo Federal, a estatal analisada tem como premissa promover qualidade de vida e desenvolvimento sustentável das cidades provendo transporte coletivo sobre trilhos e prestando o melhor serviço possível. Assim, para indicar onde os recursos devem ser empregados eficientemente é necessária a aferição e seleção dos indicadores chave de desempenho (Key Performance Indicator [KPI], em inglês).

De acordo com Uchoa (2013), para ser considerada como indicador uma variável deve ser crítica e demandar controle para que se mantenha em níveis predeterminados. Entretanto, é fundamental monitorar aqueles com maior valor agregado e implementados mais facilmente. Tais indicadores são os mais críticos para o sucesso da corporação, focados na performance organizacional (PARMENTER, 2015).

Monitorar indicadores é uma tarefa desafiadora e utilizá-los inadequadamente pode causar um impacto neutro ou até mesmo negativo na produtividade (ANGOSS, 2011). Conforme Peral *et al.* (2017), a falta dessa clareza leva a escolha de indicadores por meio da intuição, podendo resultar em consequências desastrosas. Assim, deve-se buscar tecnicamente quais indicadores devem ser observados. Chen *et al*. (2014) declaram que os resultados de uma análise só serão efetivamente utilizados se forem apresentados de maneira amigável. Por isso, para Parmenter (2015), as empresas deveriam monitorar até 10 KPIs para uma operação bem-sucedida.

O presente trabalho se propõe a aplicar a metodologia de Análise Fatorial por Componentes Principais (em inglês, *Principal Component Analysis* [PCA]), com o objetivo sugerir um subconjunto da base de dados original que contenha a maior parte da variância (informação) e possa representar suficientemente bem os dados originais, facilitando a visualização e a relação entre os registros obtidos permitindo uma tomada de decisão baseada em dados. A PCA consiste em uma técnica multivariada exploratória de análise de dados, sem caráter preditivo, onde o objetivo é conseguir expressar a informação contida nas variáveis originais em um conjunto reduzido de fatores (HAIR JR *et al*., 2009) não correlacionados entre si ao

longo dos quais a variância nos dados é máxima (ou seja, ortogonais), de forma que a dimensionalidade do conjunto de dados possa ser reduzida sem perda significativa de informação por meio de combinações lineares das variáveis originais (KASSAMBARA, 2017).

2. Material e Métodos

Os dados da pesquisa foram fornecidos pela Superintendência de uma Estatal Federal de transporte coletivo sobre trilhos, responsável pela gestão do transporte ferroviário que interliga quatro municípios da Região Metropolitana de João Pessoa/PB, ao longo de 30km. Os dados utilizados foram obtidos junto a Coordenação de Operação ao longo de 59 meses entre os anos de 2010 até 2015, no Centro de Controle Operacional [CCO], setor responsável por coletar e registrar as informações.

De acordo com o manual interno da companhia, chamado "INDICADORES DE DESEMPENHO", a empresa monitora 46 indicadores, descritos na Tabela 1. Foram utilizados 33 indicadores (a saber, os indicadores 1-9, 14, 16-29, 31-39). Os demais não foram incluídos por não se enquadrarem no propósito deste trabalho ou por ausência de dados. Os arquivos .xlsx obtidos passaram por um pré-processamento utilizando a linguagem de programação Pyhton e em seguida, todo o resto do trabalho foi realizado na linguagem de programação R.

A estrutura do conjunto de dados utilizado apresenta os indicadores originais em colunas (variáveis) e os meses das coletas dos dados em linhas (indivíduos). Deste conjunto de dados foi obtida a matriz de correlações ρ, conforme demostrado no sist. (1)

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \quad (1)$$

e neste contexto, ρ se refere às correlações de Pearson entre cada par de variáveis; quando o valor da correlação é próximo de ±1 isto indica que o mesmo fator pode ser extraído deste par. Por outro lado, numa correlação próxima de zero mais de um fator pode ser extraído a partir do par de variáveis.

Tabela 1 - Lista dos Indicadores utilizados neste trabalho

Código do Indicador	Nome do Indicador
Ind_01	Indicador de Regularidade (%)
Ind_02	Indicador de Pontualidade (%)
Ind_03	Taxa de Ocupação (passageiros/m²)
Ind_04	Indicador de Segurança do Usuário (acidentados por milhão de passageiros)
Ind_05	Indicador de Ocorrências Policiais (%)
Ind_06	Indicador de Acidentes com Empregados (%)
Ind_07	Indicador de Segurança do Sistema (Acidentes por 100.000 km)
Ind_08	Taxa de Cobertura Plena (%)
Ind_09	Passageiro Equivalente
Ind_10	Indicador de Nível de Integração Operacional entre Modos de Transporte (%)
Ind_11	Indicador de Nível de Integração Tarifária entre Modos de Transporte (%)
Ind_12	Índice de Operação no Pico do Pico
Ind_13	Índice de Operação na Hora de Pico - TOP
Ind_14	Índice de Disponibilidade de Locomotivas / Carros (%)
Ind_15	Índice de Atendimento à Demanda nos Picos
Ind_16	Índice de Imobilização de Locomotivas / Carros (%)
Ind_17	Tarifa Efetiva Média (R$)
Ind_18	Carro.km (km rodado por carro da frota)
Ind_19	Custo Unitário (R$)
Ind_20	Índice de Custo Pessoal (%)
Ind_21	Índice de Confiabilidade do Sistema Operacional – MKBF
Ind_22	Índice de Confiabilidade do Material Rodante
Ind_23	Índice de Quilometragem média entre Retiradas - MKBR
Ind_24	Índice de Avarias
Ind_25	Tempo médio de Reparo - MTTR
Ind_26	Tempo médio de liberação do Material Rodante – TML
Ind_27	Tempo médio da Operação – TMO
Ind_28	Índice de Viagens Canceladas Atribuídas aos Equipamentos – IVCAE
Ind_29	Índice de Viagens Canceladas Atribuídas aos Equipamentos no Pico
Ind_30	Índice de Compras e Contratações pela Manutenção
Ind_31	Custo de Manut. do Material Rodante por Carro Quilômetro – CMRCQ (R$)
Ind_32	Custo de Manut. das Instalações Fixas por km de Via Mantido – CMIFQV (R$)
Ind_33	Custo de Manutenção por Passageiro Quilometro – CMNPQ (R$)
Ind_34	Produtividade da Energia – IPE (km/1.000 l)
Ind_35	Produtividade do Pessoal (passageiro.km / força de trabalho)
Ind_36	Receita Total / Força de Trabalho (R$)
Ind_37	Efetivo de Maquinistas / Frota Disponível
Ind_38	Efetivo de Maquinistas / Frota Existente
Ind_39	Pessoal Alocado em Estação / Linha de Bloqueio
Ind_40	Indicador de causa trabalhista
Ind_41	Indicador de Absenteísmo – IAB
Ind_42	Taxa de Frequência de Acidentes de Trabalho
Ind_43	Indicador do orçamento / liberado (R$)
Ind_44	Indicador do orçamento liberado / executado (R$)
Ind_45	Indicador Tarifário (%)
Ind_46	Indicador da Receita Extra-operacional – IREX (%)

Fonte: Dados originais da pesquisa

De acordo com Hair Jr *et al*. (2019), a aplicação da PCA mostra-se inadequada se a matriz ρ tiver poucos valores superiores a 0,3 e para realizar essa verificação aplica-se a estatística Kaiser-Meyer-Olkin [KMO], que mede o grau de intercorrelações entre as variáveis, variando entre os valores entre zero e um, onde um indica que cada variável é plenamente predita pelas outras variáveis. Outro teste utilizado é o de esfericidade de Bartlett, que mede a presença de correlações entre as variáveis por meio da comparação da matriz ρ com uma matriz identidade. É possível aplicar a PCA se os valores fora da diagonal principal forem estatisticamente diferentes de zero (valor-p < 0,05). Com base nos resultados destes testes é possível avaliar se a PCA é adequada à base de dados (FÁVERO; BELFIORE, 2017).

Se a aplicação da PCA é apropriada, o próximo passo para obtenção dos Fatores Principais, é calcular os escores fatoriais, que são os coeficientes que relacionam os fatores com as variáveis originais, obtidos a partir dos autovalores e autovetores da matriz ρ, e correspondem aos parâmetros de um modelo de regressão linear múltipla tendo as variáveis originais representando as variáveis explicativas do modelo e os Fatores como as variáveis dependentes (FÁVERO; BELFIORE, 2017), conforme o sist. (2)

$$\begin{aligned} F_1 &= s_{11} \cdot X_1 + s_{21} \cdot X_2 + \cdots + s_{k1} \cdot X_k \\ F_2 &= s_{12} \cdot X_1 + s_{22} \cdot X_2 + \cdots + s_{k2} \cdot X_k \\ &\vdots \\ F_k &= s_{1k} \cdot X_1 + s_{2k} \cdot X_2 + \cdots + s_{kk} \cdot X_k \end{aligned} \tag{2}$$

com os Fatores Principais representados pelas letras $\boldsymbol{F}$ ($F_1, F_2, \cdots, F_k$) e seus índices referentes a cada fator derivado das variáveis principais, variando de 1 até a quantidade total de varáveis originais, $\boldsymbol{k}$; $\boldsymbol{X}$ representam as variáveis métricas originais; os $\boldsymbol{s}$ são os escores fatoriais.

Os autovalores medem a variância retida por cada Componente Principal e apresentam a característica de serem sempre maiores que o componente subsequente ($\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2$) (FÁVERO; BELFIORE, 2017; KASSAMBARA, 2017). Assim, cada Componente Principal é associado ao seu respectivo autovalor. Os $\boldsymbol{k}$ autovalores λ^2 são obtidos a partir das raízes do polinômio característico apresentado na exp. (3):

$$p(\lambda^2) = det(\lambda^2 \cdot I - \rho) \tag{3}$$

onde $\boldsymbol{I}$ é uma matriz identidade de mesma dimensão de ρ. Desenvolvendo a eq. (1) na eq. (3) e igualando a eq. (3) a zero para calcular suas raízes, obtêm-se a eq. (4):

$$\begin{vmatrix} \lambda^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda^2 - 1 \end{vmatrix} = 0 \quad (4)$$

A soma total das variâncias explicadas por cada Fator é igual à soma dos autovalores deste conjunto de variáveis ($\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2 = k$). Para obtenção dos autovetores (v) da matriz ρ a partir dos autovalores (λ^2) é necessário resolver o sist. (5) para cada autovalor. Para determinar os autovetores $v_{1k}, v_{2k}, \cdots, v_{kk}$ a partir do autovalor $\lambda_k^2 (k = 1,2, \ldots, quantidade\ de\ variáveis)$ temos:

$$\begin{pmatrix} \lambda_k^2 - 1 & -\rho_{12} & \cdots & -\rho_{1k} \\ -\rho_{21} & \lambda_k^2 - 1 & \cdots & -\rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{k1} & -\rho_{k2} & \cdots & \lambda_k^2 - 1 \end{pmatrix} \cdot \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{kk} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (5)$$

que resulta no sist. (6):

$$\begin{cases} (\lambda_k^2 - 1) * v_{1k} - \rho_{12} * v_{2k} \cdots - \rho_{1k} * v_{kk} = 0 \\ -\rho_{21} * v_{1k} + (\lambda_k^2 - 1) * v_{2k} \cdots - \rho_{2k} * v_{kk} = 0 \\ \vdots \\ -\rho_{k1} * v_{1k} - \rho_{k2} * v_{2k} \cdots + (\lambda_k^2 - 1) * v_{kk} = 0 \end{cases} \quad (6)$$

É possível obter os escores fatoriais s de cada Fator Principal F a partir dos autovetores (v) e os autovalores (λ^2). O *k*-ésimo escore pode é calculado desenvolvendo a exp. (7):

$$\boldsymbol{S}_k = \begin{pmatrix} s_{1k} \\ s_{2k} \\ \vdots \\ s_{kk} \end{pmatrix} = \begin{pmatrix} \frac{v_{1k}}{\sqrt{\lambda_k^2}} \\ \frac{v_{2k}}{\sqrt{\lambda_k^2}} \\ \vdots \\ \frac{v_{kk}}{\sqrt{\lambda_k^2}} \end{pmatrix} \quad (7)$$

Após a obtenção dos escores fatoriais utiliza-se o critério de Kaiser para selecionar a quantidade mínima de Fatores que representem adequadamente as variáveis originais, no qual os fatores selecionados possuem autovalores maiores que um ou próximos deste valor (FÁVERO; BELFIORE, 2017; HAIR JR *et al.*, 2009; KASSAMBARA, 2017), ou seja, retendo o equivalente à variância de pelo menos uma variável.

Em sequência, obtém-se os valores das cargas fatoriais que são as correlações de Pearson entre as variáveis originais e os Fatores obtidos (FÁVERO; BELFIORE, 2017). Os valores destas correlações são mostrados na matriz de correlações apresentada na Tabela 2.

Tabela 2. Cargas fatoriais entre variáveis originais e fatores

Variável	F_1	F_2	⋯	F_k
X_1	c_{11}	c_{12}	⋯	c_{1k}
X_2	c_{21}	c_{22}	⋯	c_{2k}
⋮	⋮	⋮	⋱	⋮
X_k	c_{k1}	c_{k2}	⋯	c_{kk}

Fonte: FÁVERO; BELFIORE (2017)

As cargas fatoriais são determinantes na interpretação do quanto cada variável contribui para a formação dos Fatores Principais. Se os quadrados das cargas fatoriais em cada linha forem somados o resultado será igual a 1 ($c_{k1}^2 + c_{k2}^2 + \cdots + c_{kk}^2 = 1$), porém depois da aplicação do critério de Kaiser isso não ocorre, sendo este resultado chamado de comunalidade, conforme apresentado no sistema (8), e representa a variância compartilhada de cada variável com todos os fatores remanescentes na análise (FÁVERO; BELFIORE, 2017).

$$\begin{aligned} c_{11}^2 + c_{12}^2 + \cdots &= comunalidade\ de\ X_1 \\ &\vdots \\ c_{k1}^2 + c_{k2}^2 + \cdots &= comunalidade\ de\ X_k \end{aligned} \tag{8}$$

onde c_{ij} se refere à carga fatorial entre a variável $\boldsymbol{i}$ e o fator $\boldsymbol{j}$. Do resultado da análise de comunalidade é possível estimar se há alguma variável que não colabora significativamente para a extração dos Fatores e avaliar a sua remoção da PCA.

Um típico resultado de uma PCA consiste em encontrar os autovalores (fornecem informações sobre a variabilidade dos dados), os escores fatoriais (fornecem informações sobre a estrutura das observações) e as cargas fatoriais (correlações de Pearson entre as variáveis e os Fatores extraídos) (FÁVERO; BELFIORE, 2017).

3. Resultados e Discussão

A estrutura básica dos dados utilizados pode ser observada na Tabela 3 e com base nela destacam-se alguns pontos. De acordo com o manual, os 3 primeiros indicadores se referem aos Indicadores de Nível de Serviço Oferecido e se mostram relativamente constantes ao longo do período observado. Do Ind_04 ao Ind_07 são os Indicadores de Segurança e destaca-se o fato de que o Ind_04 e o Ind_05 apresentam ocorrências pontuais enquanto o Ind_06 e Ind_07 mostram ocorrências mais frequentes. O Ind_08 e o Ind_09 tratam da matéria de Eficácia do sistema e tiveram uma leve tendência de queda. A Eficiência é monitorada pelos Indicadores 14 a 29, de onde destacam-se o aumento do Ind_19 e o Ind_20, relativos aos custos. Os indicadores 31 a 33 tratam de Indicadores de Custo de Manutenção e apresentam

tendência de aumento possivelmente em virtude da depreciação dos veículos. Os indicadores de Produtividade (34 a 39) apresentam valores relativamente constantes para os 3 primeiros, enquanto os outros 3, ligados a pessoal, tiveram tendência de queda.

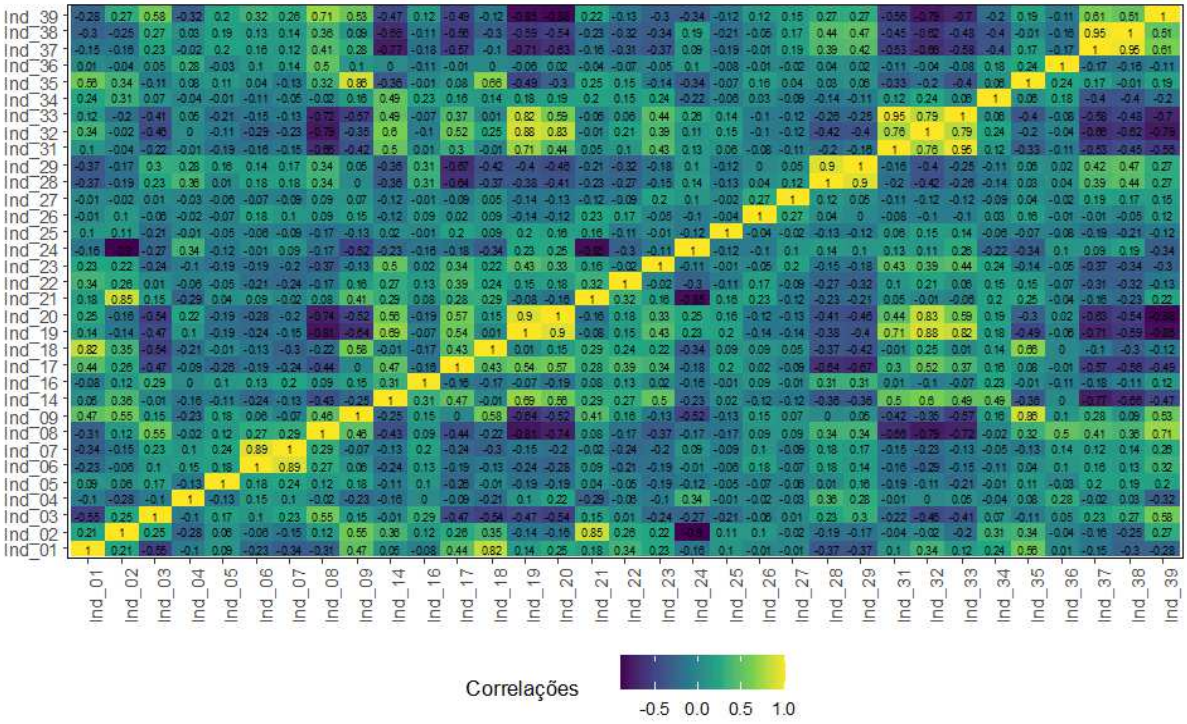
Tabela 3. Estrutura básica dos dados utilizados neste trabalho

Mês	Ind_01	Ind_02	Ind_03	…	Ind_36	Ind_37	Ind_38	Ind_39
2011-01	92,54967	80,14311	2,506537	…	388,4596	1,86	1,072488	7,4375
2011-02	90,12097	80,3132	2,408755	…	264,2645	1,765766	1,053763	7,4375
2011-03	92,2956	78,02385	1,873406	…	320,5172	1,722222	0,970194	7,4375
2011-04	98,28125	69,63434	2,195225	…	320,3037	1,730769	0,938897	7,4375
2011-05	91,55354	63,75618	2,039457	…	341,9448	1,844193	1,058537	7,25
2011-06	90,74355	68,39465	2,692584	…	530,7088	1,25	0,791457	7,5625
⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
2015-06	98,9426	86,41221	1,688851	…	661,0971	1,211632	0,838926	5,125
2015-07	98,88424	84,06206	1,74327	…	320,0883	1,167169	0,807292	5,125
2015-08	96,25	78,125	1,946331	…	312,4953	1,295987	0,900116	5,125
2015-09	98,33333	86,13251	1,716081	…	302,6714	1,209677	0,837989	5,125
2015-10	99,4152	89,16914	1,682607	…	323,6351	1,015625	0,691489	5,125
2015-11	99,84202	94,13629	1,87379	…	306,3632	1,04712	0,698487	5,125

Fonte: Resultados originais da pesquisa

Após a visualização prévia dos dados na sua forma tabular, foi elaborada a matriz de correlações ρ mostrada na Figura 1.

Figura 1 - Matriz de correlações ρ para os dados deste trabalho



Fonte: Resultados originais da pesquisa

Inicialmente percebe-se uma correlação forte entre um grupo de indicadores. São eles Ind_19, Ind_20, Ind_31, Ind_32 e Ind_33. Eles correspondem a indicadores de custo e mantém uma correlação positiva bem acentuada. Em contrapartida, este primeiro grupo apresenta correlação negativa com outro grupo, formado por Ind_37, Ind_38 e Ind_39, o que pode indicar que um efetivo maior para conduzir os veículos provoca uma diminuição nos custos de manutenção desses equipamentos (provavelmente por haver mais pessoas cuidando dos equipamentos) bem como o aumento de funcionários nas estações faz com que o zelo pelos edifícios seja maior, diminuindo assim os gastos com manutenção das instalações fixas.

No mesmo sentido destacam-se os indicadores Ind_02 e Ind_21 fortemente correlacionados, sendo este último traduzido por quilometragem média que os veículos percorrem sem falhas, que claramente irá ser refletido na pontualidade; por outro lado, esses mesmos indicadores têm uma evidente correlação negativa com o Ind_24, apontando que o acúmulo de avarias nos veículos tende a provocar atrasos e falhas.

Para concluir esta análise preliminar é possível notar ainda algumas correlações mais intuitivas, como entre o Ind_06 e o Ind_07, diretamente correlacionados entre si ou ainda a correlação entre o Ind_08 e o Ind_19, inversamente correlacionados.

Uma vez obtida a matriz ρ foi aplicada a função KMO(r = rho) do pacote psych (*version* 2.2.5) da linguagem R que avalia a tabela fornecida e retorna um valor cujo resultado pode ser enquadrado em uma das condições apresentadas na Tabela 4. Para a base de dados do estudo, o resultado obtido foi de 0,631066 que a enquadra como razoável para aplicação da PCA.

Tabela 4. Relação entre a estatística KMO e a adequação global da análise fatorial

Estatística KMO	Adequação Global da Análise Fatorial
Entre 1,00 e 0,90	Muito boa
Entre 0,90 e 0,80	Boa
Entre 0,80 e 0,70	Média
Entre 0,70 e 0,60	Razoável
Entre 0,60 e 0,50	Má
Menor do que 0,50	Inaceitável

Fonte: FÁVERO; BELFIORE (2017)

Em seguida foi aplicada a função cortest.bartlett(R = rho) utilizada para realizar o teste de Bartlett e fornece como saída o valor da estatística $\chi^2_{Bartlett}$, o valor-p e os graus de liberdade. Para esta base de dados o valor-p foi igual a zero indicando que as correlações de Pearson são estatisticamente diferentes de zero e a extração de Fatores se mostra adequada.

Uma vez que foi confirmada a adequação da base de dados, a função PCA() [pacote *FactoMineR*] (LÊ *et al.*, 2008) foi utilizada gerando um objeto contendo autovalores e algumas outras listas e matrizes. Para a visualização dos resultados obtidos foram utilizadas funções que fazem parte do [pacote *factoextra*] (KASSAMBARA; MUNDT, 2020).

Do objeto criado no passo anterior, por meio da função get_eigenvalue() foi obtida uma tabela na qual a primeira coluna apresenta o autovalor associado a cada Fator, na segunda coluna os seus percentuais com relação a variância total da base e uma coluna com a variância cumulativa, conforme observado na Tabela 5.

Tabela 5 - Autovalor, porcentagem da variância compartilhada e variância compartilhada cumulativa

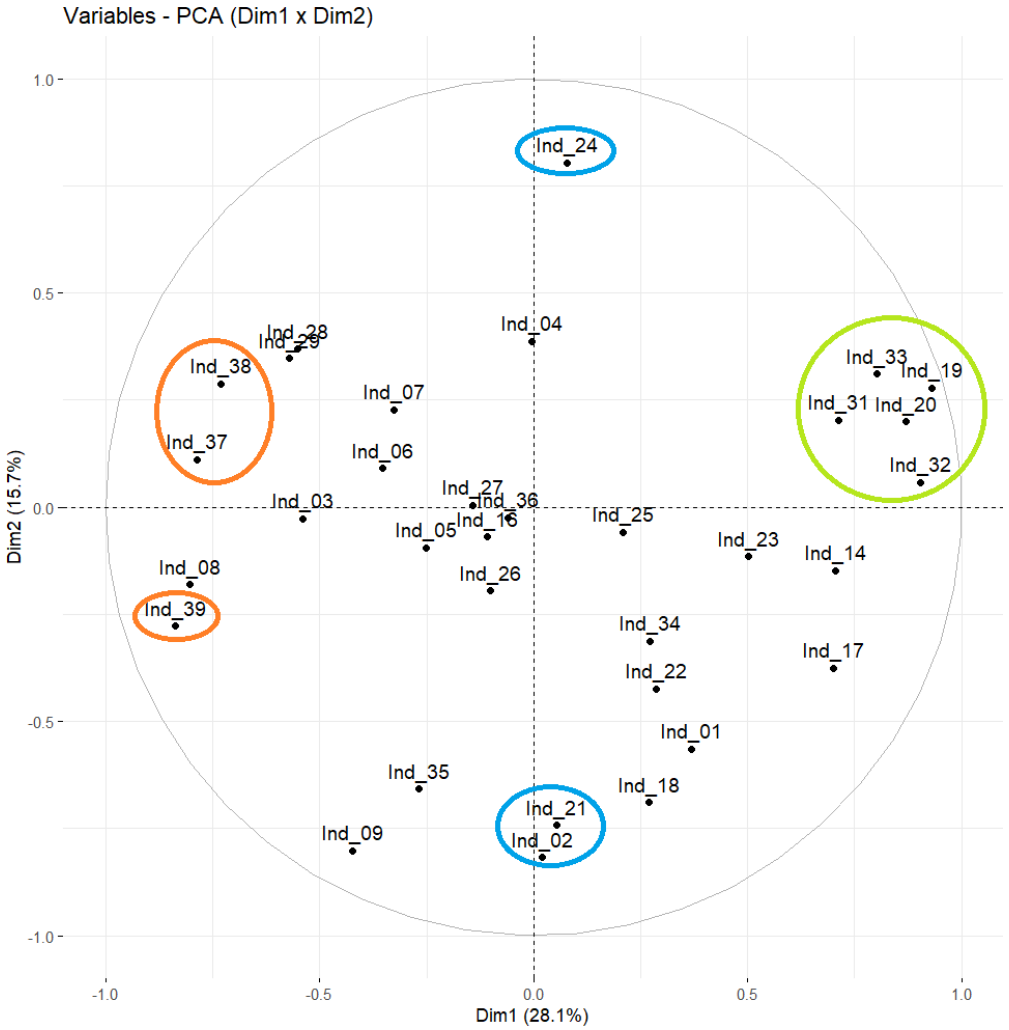
Fator	Autovalor	Variância compartilhada	Variância compartilhada cumulativa
01	9.272507309	28.098506996	28.09851
02	5.177544260	15.689528061	43.78804
03	2.980698001	9.032418185	52.82045
04	2.056675551	6.232350155	59.05280
05	1.825284416	5.531164897	64.58397
06	1.580604346	4.789710139	69.37368
07	1.393032775	4.221311438	73.59499
08	1.233077648	3.736598934	77.33159
09	1.207039512	3.657695492	80.98928
10	0.993470041	3.010515277	83.99980
⋮	⋮	⋮	⋮
31	0.007096791	0.021505428	99.97965
32	0.004344612	0.013165490	99.99281
33	0.002371572	0.007186582	100.00000

Fonte: Resultados originais da pesquisa

Como é possível notar, os 10 primeiros fatores atendem ao critério de Kaiser, os quais conseguem reter praticamente 84% da variância total dos dados. Conforme Hair Jr *et al*. (2019), uma porcentagem acima de 60% de variância explicada já é suficiente para obtenção de uma solução adequada.

Obtidos os fatores é possível mostrar a correlação entre as variáveis originais de uma maneira alternativa, plotando a distribuição dos indicadores num plano cartesiano formado por pares de Fatores, conhecido por *loading plot*, conforme apresentado na Figura 3, gerado pela função fviz_pca_var(). Para esta visualização foram utilizados como eixos os dois fatores com maior variância (neste gráfico os fatores são chamados de "Dim" e ao lado de cada fator está indicada a sua variância compartilhada).

Figura 3 - Círculo de correlação para os dois primeiros Fatores



Fonte: Resultados originais da pesquisa

A Figura 3 pode ser interpretada da seguinte maneira: variáveis positivamente correlacionadas estão próximas entre si enquanto negativamente correlacionadas estão em lados opostos; quanto mais distante da origem, mais bem representada esta variável está. Assim, ratifica-se a análise feita a respeito da relação entre alguns indicadores. Inicialmente é possível notar a proximidade entre os Ind_19, Ind_20, Ind_31, Ind_32 e Ind_33, (círculo verde), representando uma forte correlação positiva enquanto os Ind_37, Ind_38 e Ind_39 (elipses na cor laranja) estão do lado oposto. Por fim também foi dado destaque (em azul) aos Ind_02 e Ind_21 positivamente correlacionados entre si e em oposição ao Ind_24. Esse gráfico pode ser gerado para cada par de Fatores, entretanto a explicabilidade dos Fatores diminui à medida que os autovalores diminuem.

Após aplicação do critério de Kaiser foram obtidas as comunalidades. A Tabela 6 foi elaborada para apresentar as comunalidades para os dados deste estudo.

Tabela 6 - Cargas Fatoriais dos Indicadores com as 10 maiores Comunalidades

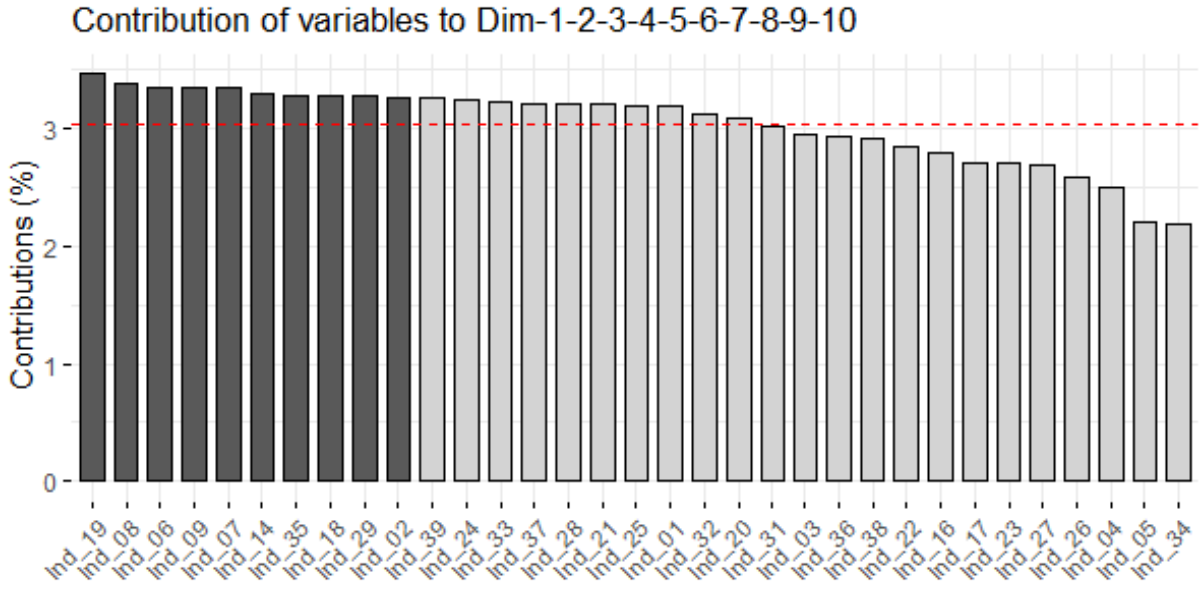
Ind	PC01	PC02	PC03	PC04	PC05	PC06	PC07	PC08	PC09	PC10	Comunalidade
19	0,930	0,278	-0,107	-0,012	-0,042	-0,050	0,032	0,019	0,000	0,022	0,960
08	-0,804	-0,180	-0,147	-0,261	-0,023	0,345	-0,142	0,045	0,151	0,001	0,934
06	-0,353	0,091	-0,240	-0,455	-0,576	-0,428	-0,058	-0,044	0,045	-0,084	0,926
09	-0,423	-0,801	0,185	-0,155	0,140	-0,080	0,096	-0,084	0,061	-0,017	0,926
07	-0,327	0,228	-0,361	-0,436	-0,565	-0,311	0,020	-0,137	0,101	0,009	0,926
14	0,705	-0,148	-0,574	0,013	0,122	0,117	0,023	-0,120	0,095	0,096	0,910
35	-0,269	-0,658	0,433	-0,392	0,183	-0,095	0,078	0,051	-0,022	-0,092	0,907
18	0,270	-0,689	0,533	-0,113	0,010	-0,215	0,038	-0,032	0,112	0,009	0,907
29	-0,571	0,347	-0,161	-0,143	0,519	-0,300	0,138	0,176	-0,018	-0,053	0,905
02	0,020	-0,818	-0,378	0,128	0,073	-0,049	-0,036	0,222	0,073	-0,093	0,902

Fonte: Resultados originais da pesquisa

A função fviz_contrib() foi utilizada para gerar a Figura 4, que apresenta como as variáveis originais contribuem para a formação dos 10 primeiros Componentes Principais. Conforme pode ser observado, é possível chegar às seguintes observações:

- As barras em cinza escuro correspondem às 10 variáveis com maior contribuição para geração dos 10 primeiros Componentes Principais extraídos da base de dados;
- A linha tracejada em vermelho indica a contribuição média esperada (percentual). Numa distribuição uniforme o resultado seria de 1/33×100≅3,03%.

Figura 4 - Contribuição das 33 variáveis para construção dos 10 primeiros Fatores



Fonte: Resultados originais da pesquisa

Com os resultados obtidos das análises apresentadas é possível confirmar que algumas variáveis são bastante correlacionadas entre si, como é perceptível na matriz ρ. Esse tipo de

relação entre as variáveis mostra que há muita redundância na base de dados que a tornam demasiadamente complicada e custosa de analisar sem um preparo prévio adequado (PERAL *et al*, 2017), uma vez que o volume de dados tende a aumentar e dificultar o acompanhamento apropriado dos indicadores. Assim, na era do *Big Data*, é necessário para o desenvolvimento empresarial ter a capacidade de acessar, processar e transformar os dados em ideias valiosas capazes de trazer resultados como agilidade, inovação, e desempenho competitivo (MIKALEF *et al*., 2020).

Realizando o mapeamento dos indicadores utilizando a Tabela 1, foram obtidos os dez primeiros indicadores, conforme a mesma ordem apresentada tanto na Figura 4 como na Tabela 6:

1. Ind_19. Custo Unitário (R$)
2. Ind_08. Taxa de Cobertura Plena (%)
3. Ind_06. Indicador de Acidentes com Empregados (%)
4. Ind_09. Passageiro Equivalente
5. Ind_07. Indicador de Segurança do Sistema (acidentes por composição.km)
6. Ind_14. Índice de Disponibilidade de TUEs / Locomotivas / Carros (%)
7. Ind_35. Produtividade do Pessoal (passageiro.km / força de trabalho)
8. Ind_18. Carro.km (km rodado por carro da frota)
9. Ind_29. Índice de Viagens Canceladas Atribuídas aos Equipamentos no Pico
10. Ind_02. Indicador de Pontualidade (%)

Verificando os itens dessa lista percebemos que os cinco mais relevantes são voltados para as questões de custo (Ind_19, Ind_08 e Ind_09) e segurança (Ind_06, Ind_07). Daí, já é possível indicar para os gestores a necessidade de direcionar os esforços e recursos para essas duas áreas, com a expectativa de que, baseados nos dados entre os anos 2010 e 2015, a empresa possa melhore a gestão ao manter o foco nesses indicadores. Dos outros cinco restantes, três estão ligados a manutenção (Ind_14, Ind_18, Ind_29), outro grande pilar para manter a empresa estudada em funcionamento eficiente.

O resultado desta análise não deve se resumir apenas na mera indicação de investimento nas áreas de custo, segurança ou manutenção, mas também dar o poder de ajudar na tomada de decisão, sendo cada vez mais fundamentadas em análises consistentes de dados, ou seja, em ciência e pesquisa. Isso se torna ainda mais importante quando se refere à aplicação de verba pública para atendimento aos cidadãos.

Para melhorar os resultados das análises obtidas neste trabalho poderia ser seguida a abordagem sugerida por Rodríguez-Rodríguez *et al* (2009) que sugere a aplicação de modelos de regressão por mínimos quadrados parciais de forma que sejam identificadas as relações entre os indicadores e como cada um exerce influência sobre os demais.

Com o mesmo intuito de tornar o resultado do trabalho ainda mais preciso, Halawa *et al* (2021) utilizam técnicas de agrupamento para agregar variáveis de acordo com o seu comportamento com o objetivo de facilitar a análise em razão da diminuição da quantidade de variáveis por exclusão daquelas que não estariam em um grupo relevante.

É importante salientar que a PCA deve ser executada por completo sempre que a base de dados sofrer qualquer alteração como a inclusão de um novo mês ou um novo indicador, pois os parâmetros obtidos para esta base de dados são em função dos dados disponíveis (FÁVERO; BELFIORE, 2017). Qualquer alteração irá modificar os autovalores e os autovetores obtidos, e consequentemente todos os demais resultados, os quais precisarão ser recalculados e certamente levarão a respostas diferentes das alcançadas neste trabalho.

4. Conclusão

Por meio do uso da PCA foi possível a obtenção dos indicadores Ind_19, Ind_08 e Ind_09, relativos aos custos; os indicadores Ind_06 e Ind_07 relativos à segurança; os indicadores Ind_14, Ind_18, Ind_29, relativos à manutenção; além dos Ind_35 e Ind_02 como sendo os indicadores que mais causam impacto no desempenho operacional da empresa estudada dentro do período observado.

Dessa forma, foi possível desenvolver meios para a elaboração de uma ferramenta de visualização para que os gestores acompanhem esses indicadores e possam atuar de maneira baseados em dados científicos e métodos comprovadamente eficazes, de forma que a Companhia continue firme como referência no mercado de transporte de passageiros.

Para estudos futuros poderiam ser aplicadas técnicas de aprendizado de máquina não supervisionada para agrupar as variáveis conforme as suas características e facilitar a alocação de recursos por área específica. Além disto, é possível a aplicação de métodos que possam relacionar as variáveis e apontar onde cada uma exerce mais influência por meio de técnicas de regressão, de forma a conseguir otimizar o controle de cada um dos indicadores.

Referências

ANGOSS. 2011. Key performance indicators, six sigma, and data mining. White Paper. Disponível em: <https://www.yumpu.com/en/document/read/49399047/key-performance-indicators six-sigma-and-data-mining-angoss->. Acesso em: 15 abr. 2022.

ANPTrilhos. 2018. Mobilidade urbana sobre trilhos na ótica dos grandes formadores de opinião. Associação Nacional dos Transportadores de Passageiros sobre Trilhos [ANPTrilhos]. 182 p. Brasília, DF, Brasil. Disponível em: <https://anptrilhos.org.br/wp-content/uploads/2018/08/ANPTrilhos_livro_Formadores_Opiniao_web.pdf>. Acesso em: 27 Set. 2022.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. 2014. Big Data: A Survey. Mobile networks and applications, 19(2), pp.171–209.

CLARK, André. 2018. A importância dos trilhos para a mobilidade urbana brasileira. In: Mobilidade urbana sobre trilhos na ótica dos grandes formadores de opinião. Associação Nacional dos Transportadores de Passageiros sobre Trilhos [ANPTrilhos]. pp. 28-30. Brasília, DF, Brasil. Disponível em: <https://anptrilhos.org.br/wp-content/uploads/2018/08/ANPTrilhos_livro_Formadores_Opiniao_web.pdf>. Acesso em: 27 Set. 2022.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata. Rio de Janeiro: GEN LTC.

HAIR Jr, Joseph F.; BLACK, William C.; BABIN, Barry J.; ANDERSON, Rolph E.; TATHAM, Ronald L. 2009. Análise multivariada de dados. 6 ed. Porto Alegre: Bookman.

HALAWA, Mohamed Soliman; DÍAZ REDONDO, Rebeca P.; FERNÀNDEZ VILAS, Ana. 2021. KPIs-Based Clustering and Visualization of HPC Jobs: A Feature Reduction Approach. IEEE Access, vol. 9, pp. 25522-25543.

HOBBS, Jason; BAIMA, Carollina; SEABRA, Renata; IDOM Consulting. 2021. Desenvolvimento orientado ao transporte: Como criar cidades mais compactas, conectadas e coordenadas. Monografia. Banco Interamericano de Desenvolvimento. Brasil. Disponível em: <https://publications.iadb.org/publications/portuguese/document/Desenvolvimento-orientado-ao-transporte-Como-criar-cidades-mais-compactas-conectadas-e-coordenadas.pdf>. Acesso em: 11 Abr. 2022.

KASSAMBARA, Alboukadel. 2017. Practical Guide to Principal Component Methods in R. STHDA - Statistical tools for high-throughput data analysis. 1 ed. United States: Stdha.com.

KASSAMBARA, Alboukadel; MUNDT, Fabian. 2020. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7.

LÊ, Sébastien; JOSSE, Julie; HUSSON, François. 2008. "FactoMineR: A Package for Multivariate Analysis." Journal of Statistical Software, 25(1), 1–18.

MIKALEF, Patrick; PAPPAS, Ilias; KROGSTIE, John; PAVLOU, Paul. 2020. Big data and business analytics: A research agenda for realizing business value. Information & Management, 57 (1), Article 103237.

PARMENTER, David. 2020. Key Performance Indicators: Developing, Implementing, and Using Winning KPIs. 4 ed. Hoboken, New Jersey: Wiley.

PERAL, Jesús; MATÉ, Alejandro; MARCO, Manuel. 2017. Application of data mining techniques to identify relevant key performance indicators. Computer Standards & Interfaces 54(2): 76-85.

RODRÍGUEZ-RODRÍGUEZ, Raul; ALFARO-SAIZ, Juan Jose; ORTIZ BAS, Angel. 2009. Quantitative relationships between key performance indicators for supporting decision-making processes. Computers in Industry, 60 (2). Pp. 104-113.

UCHOA, Carlos Eduardo. 2013. Elaboração de indicadores de desempenho institucional. Escola Nacional de Administração Pública [ENAP]. Brasília, Brasil. Disponível em: <https://repositorio.enap.gov.br/bitstream/1/2403/1/Elabora%C3%A7%C3%A3o%20de%20indicadores%20de%20desempenho_apostila%20exerc%C3%ADcios.pdf>. Acesso em: 11 Abr. 2022.

UNITED NATIONS. 2019. World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420). Department of Economic and Social Affairs, Population Division. New York, USA.