Universidade Estadual da Paraíba

Centro de Ciências e Tecnologia

Departamento de Estatística

Wanessa Isthéwany de Albuquerque Wanderley

Análise de componentes principais aplicados à ocorrência de acidentes de trânsito em Campina Grande - PB

Campina Grande - PB

Agosto de 2017

Wanessa Isthéwany de Albuquerque Wanderley

Análise de componentes principais aplicados à ocorrência de acidentes de trânsito em Campina Grande - PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Profª. Drª. Ana Patrícia Bastos Peixoto

Coorientador: Profª. Drª. Maria Joseane Cruz da Silva

Campina Grande - PB

Agosto de 2017

W245a Wanderley, Wanessa Isthéwany de Albuquerque.
Análise de componentes principais aplicados à ocorrência de acidentes de trânsito em Campina Grande - PB [manuscrito] / Wanessa Isthéwany de Albuquerque Wanderley. - 2017.
39 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2017.
"Orientação: Profa. Dra. Ana Patrícia Bastos Peixoto, Departamento de Estatística".

1. Acidentes de trânsito. 2. Análise gráfica. 3. Análise multivariada. I. Título.

21. ed. CDD 519.53

Wanessa Isthéwany de Albuquerque Wanderley

Análise de componentes principais aplicados à ocorrência de acidentes de trânsito em Campina Grande - PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 09 de Agosto de 2017 .

BANCA EXAMINADORA

Ana Patrícia Bastos Peixoto

Profa. Dra. Ana Patricia Bastos Peixoto
Universidade Estadual da Paraíba

Ricardo Alves de Olinda

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Mª das Vitórias A. Serafim

Profa. Msc. Maria das Vitórias Alexandre Serafim
Universidade Estadual da Paraíba

A minha grandiosa mãe Ismênia, meu esposo Felipe e a meus irmãos Thales, Thayse e Wagner, familiares, enfim, dedico a todos que estiveram presentes por todo o tempo, na conclusão desta etapa tão importante de minha vida.

Agradecimentos

Primeiramente agradeço a Deus, que esteve sempre ao meu lado me guiando e me iluminando em sua divina misericórdia para que eu realizasse mais uma etapa na minha vida.

Agradeço a minha mãe Ismênia, pelo seu exemplo de vida, de mãe, mulher e de amor, por não medir esforços para querer sempre o melhor para seus filhos, por estar sempre ao meu lado me apoiando, incentivando nos momentos mais difíceis.

Aos meus irmãos Thales, Thayse e Wagner, cunhada Jussara, minhas sobrinhas Júlia e Nycole, por me fazerem uma pessoa mais feliz.

Ao meu esposo Felipe, que me ajudou bastante, pelo seu apoio, dedicação, paciência e por estar sempre ao meu lado e me fazendo uma pessoa mais feliz do que já sou. Não poderia deixar de agradecer a família Barros, pela paciência e apoio.

A minha avó Lúcia, por transmitir sua experiência de vida e a todos os familiares que estiveram comigo nessa etapa da minha vida. Agradeço a meu pai e a minha família paterna.

Pela orientação das professoras Ana Patrícia e Maria Joseane, pelo partilhamento do conhecimento, apoio e confiança tornando possível a realização desse trabalho.

Aos professores que fazem parte da banca Ricardo Alves e Maria das Vitorias e a aos professores Thiago, Kleber, Gustavo, Juarez e Gisely e a todo o corpo docente do Departamento de Estatística pelo grande conhecimento repassado, ao Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, além de todo o quadro de servidores técnicos administrativos.

Aos meus amigos Aline, Arnete, Damião, em especial a Sônia pelo apoio, ajuda nos momentos de grandes dificuldades e a todo os colegas pelos momentos agradáveis dentro e fora da Universidade.

Agradeço a toda equipe que trabalha no Serviço de Atendimento Móvel de Urgência (SAMU) em especial a Fátima, Paulo e Deoclécio, pela gentiliza e disposição de fornecer os dados.

Agradeço a Superintendência de Trânsito e Transportes Públicos, pelo apoio, contribuição na finalização de uma etapa na minha vida.

Em fim a todos que acreditaram e contribuíram de alguma forma para este trabalho. Meu muito Obrigado!!!!!

"Sonhos determinam o que você quer.
Ação determina o que você conquista."
(Aldo Novak)

Resumo

O número da frota veicular no estado da Paraíba tem aumentado muito nos últimos anos, com mais fluxos de veículos nas vias, há forte indícios de mais acidentes. Diante deste problema, o objetivo deste trabalho é verificar, caracterizar e analisar os acidentes na cidade de Campina Grande-PB. Para esta finalidade, existem vários métodos que podem avaliar e estudar o comportamento dos acidentes. Uma delas é a abordagem multivariada, a qual tem como objetivo avaliar simultaneamente várias variáveis relacionadas entre si, dentre as diversas técnicas multivariadas existe a Análise de Componentes Principais. Com a análise ACP o pesquisador pode simplificar e redimensionar as variáveis num banco de dados que o explique sem perda de informação. Os dados deste trabalho foram coletados do Serviço de Atendimento Móvel de Urgência, no primeiro semestre de 2014, referentes aos acidentes de trânsito com vítimas. Em que foi utilizado a estatística descritiva e diante dos resultados obtidos, decidiu-se aplicar há um grupo de dados a análise multivariada, na intenção de reduzir e investigar a dependência entre variáveis.Aplicando a técnica de Componentes Principais obteve-se como resultado a representação de dois componentes principais explicando 95,51% dos dados originais.

Palavras-chaves:Acidentes de trânsito; Análise Gráfica; Componentes principais

Abstract

The number of traffic accidents has increased in recent years. This fact has become a common problem throughout Brazil. Faced with this problem arises as the objective of the study to verify, characterize and analyse the accidents in the city of Campina Grande-PB. Faced with this problem arises as the objective of the study to characterize the accidents in the city of Campina Grande-PB. For the latter, some methods that can evaluate and study the behavior of claims cases. One is a multivariate approach, which have as a goal evaluate simultaneously, several variables related. Among the various multivariate techniques there is an analysis of the principal components that can be used to evaluate the behaviors, considering some variables of interest. In this way, the researcher can simplify and discard variables in a database that reports without loss of information. The survey data is collected from the Emergency Mobile Service in the first half of 2014. A total of 730 incident reports were analyzed, with four sets of data relating to the accident profile, accident temporal analysis, driver behavior and victim behavior. After organizing the data for a descriptive analysis, in which, given the results obtained, install, use a data group for a multivariate analysis. Such a group comprises the type categorical variable of accidents, in addition to the individuals, which are the five zones of the city, with the intention of reducing and investigating the dependence between variables. Applying a technique of the principal components we obtained as a result a representation of two main components explaining 95.51 % of the original data.

Key-words: Traffic-accidents; Graphic analysis; Principal components

Lista de ilustrações

Figura 1 – ScreePlot ilustrativo de um exemplo com sete componentes principais 23
Figura 2 – Regiões do Município de Campina Grande 24
Figura 3 – Nº de Acidentes por Regiões 27
Figura 4 – Análise das Regiões por Tipo de Acidente 27
Figura 5 – Análise das regiões por tipo de veículo 28
Figura 6 – Análise semestral por zona 28
Figura 7 – Análise das Zonas por Dia do Mês 29
Figura 8 – Análise das zonas por dia da semana 29
Figura 9 – Condutores que utilizaram ou não equipamentos de segurança e se apresentaram ou não sinais de embriagues por Zona 30
Figura 10 – Vítimas de acidentes por sexo e zona 31
Figura 11 – Condições das vítimas por zona 31
Figura 12 – Atendimento as vítimas de acidentes por zonas 32
Figura 13 – Correlação entre os tipos de colisão. 33
Figura 14 – Scree Plot de um conjunto de dados com 4 componentes principais . . 34
Figura 15 – *Biplot* $CP1 \times CP2$ de acidentes por zonas sobre o tipo dos sinistros . . 36

Lista de tabelas

Tabela 1 – Organização das n medidas nas p variáveis. 14
Tabela 2 – Componentes principais (CPs), autovalores (λ_i) e porcentagem da variância explicada e proporção acumulada (%) pelos componentes. 34
Tabela 3 – Coeficientes de ponderação e correlação dos dois primeiros componentes principais. 35

Sumário

1 INTRODUÇÃO 11

2 FUNDAMENTAÇÃO TEÓRICA 12
2.1 Marco Histórico 12
2.2 Visão geral da análise multivariada 13
2.2.1 Análise de Componentes Principais 16
2.2.2 Construção dos Componentes Principais 17
2.2.3 Determinação dos Componentes Principais 20

3 MATERIAL E MÉTODOS 24

4 APLICAÇÕES 27
4.1 Análise gráfica dos dados 27
4.2 Aplicação da análise de Componentes Principais 33

5 CONCLUSÃO 37

REFERÊNCIAS 38

1 Introdução

Conhecer as leis, compreendê-las e respeitá-las são princípios fundamentais para garantir que todas as pessoas, indistintamente, exerçam com segurança seu direito legítimo de ir e vir e de transitar. Porém, nem sempre os condutores de veículos respeitam esses princípios. De acordo com a Polícia Rodoviária Federal PRF (2015) em 2014 no Brasil houve 168.593 acidentes e 8.227 mortos só nas rodovias, representando aproximadamente 22 mortes por dia. Porém, no ano de 2014 houve uma pequena redução dos acidentes se comparado a 2013 com uma diferença de 18.105 acidentes.

Várias são as causas desses acidentes e diversos são os tipos de veículos que se envolvem nos acidentes. De acordo com o DENATRAN (2015) até setembro de 2015, as motocicletas representam no Brasil 26% do total da frota veicular, 44% na região Nordeste, no estado da Paraíba representa o mesmo percentual.No entanto em Campina Grande, a cidade possui uma frota veicular de 155.622. Dessa frota, cerca de 56.547 são motocicletas representando 36,33%.

Considerando os fatos expostos e a frota veicular do município de Campina Grande-PB que cresce anualmente e aumenta eventualmente número de acidentes, foi realizado um estudo para avaliar o comportamento desses acidentes. Uma maneira para obter e avaliar esses acidentes é utilizando a técnica multivariada que proporciona inúmeros métodos e análises que utilizam ao mesmo tempo informações de todas as variáveis respostas na explicação dos dados.

Dentre as técnicas estatística multivariadas a análise de Componentes Principais consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis da mesma dimensão. Os Componentes Principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, são independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (JOHNSON; WICHERN, 1998) e (HONGYU; SANDANIELO; OLIVEIRA, 2015).

Diante dos fatos apresentados, a análise de componentes principais foi utilizada com o objetivo de estudar e explorar a relação existente entre as estruturas multivariadas sobre a ocorrência dos acidentes de trânsitos por Zonas em relação a variável categórica tipo de acidente na cidade de Campina Grande-PB, cuja finalidade é a obtenção de um pequeno número de combinações lineares de um conjunto de variáveis, que retenham o máximo possível da informação contida nas variáveis originais e possam ser explicadas ao número reduzido de variáveis modificadas.

2 Fundamentação Teórica

O conteúdo desta seção aborda primeiramente, o Marco Histórico sobre os acidentes de trânsito no Brasil, na segunda secção uma Visão geral da análise multivariada com ênfase na análise, construção e determinação dos Componentes Principais.

2.1 Marco Histórico

Em meados do século XIX a principal forma de locomoção ainda era o próprio ato de caminhar, até que conseguiram dominar os animais e domesticá-los para serem seu meio de transporte. Em 1891 surgiu no Brasil o primeiro automóvel movido à gasolina, cujo modelo Peugeot com motor Daimler (Procedente da Europa) pertencente a Alberto Santos Dumont (1873-1932) seria a mola propulsora que transformaria o cenário das grandes cidades, o comportamento cultural da sociedade e desenvolveria a economia do país.

Vale Salientar que foi o segundo automóvel a chegar no Brasil que fez parte do que se considera o primeiro acidente automobilístico do país. Mal imaginava o grande escritor Olavo Bilac que seria a primeira vítima de acidente de trânsito no Brasil. O fato ocorreu ainda no século XIX, mais precisamente no ano de 1897 no Rio de Janeiro, na estrada da Tijuca. Olavo Bilac dirigia o Serpollet pertencente ao amigo José do Patrocínio quando perde o controle da direção, bate numa árvore e despencara num barranco (GAZIR, 1998). Não imaginavam as pessoas que admiravam aquelas invenções, que mais tarde, essas se tornariam uma das principais causas de inúmeros acidentes e mortes no país.

Diante do primeiro acidente registrado, o Poder Público e o Automóvel Clube do Brasil, começaram a se esforçar no sentido de tornar o tráfego mais seguro, direcionando as suas ações para os pedestres e para os motoristas. Autoridades Municipais de São Paulo e do Rio de Janeiro, com o intuito de disciplinar e ordenar o trânsito de veículos, em 1903, legalizaram o trânsito de automóveis, com a concessão das primeiras licenças para dirigir, sendo que em 1906 adotava-se no país o exame obrigatório para habilitar motoristas (OLIVEIRA, 1997)

Depois da revolução industrial, o automóvel particular que antes era usado por poucos, que faziam parte da elite, começou a surgir em grande massa no mundo todo. Ele transfigurou-se em objeto de consumo e posição social. Houve a necessidade de começar a implantar as placas de trânsito no Brasil. Diante de tal situação da mobilidade urbana se criou o primeiro código de trânsito brasileiro, o qual entrou em vigor pelo Decreto lei número 2.994 em 28 de janeiro de 1941. Porém não obteve sucesso durante os 8 meses de sua vigência sendo revogado no mesmo ano pelo Decreto de lei número 3.561, o que serviu para a criação do Conselho Nacional e Regional de Trânsito.

Após a segunda guerra mundial em 1945, as motocicletas transformaram-se em

veículos comuns, porém com a dificuldade de importação acabaram ficando obsoletas e mais tarde iniciou-se a produção de bicicletas no país (PR, 2006). Em 1956 Juscelino Kubitschek assumia a Presidência da República Federativa do Brasil, com seu plano de metas popularmente conhecido como "Cinquenta anos em cinco". Os aspectos mais importantes desenvolvidos no seu governo, foram: a implantação da indústria automobilística com a vinda de fábricas de automóvel para o Brasil, a construção de Brasília e a construção das rodovias ligando as regiões brasileiras (LESSA, 2005). Diante deste cenário, multiplicaram-se estradas e veículos, aumentando o fluxo dos mesmos e as pessoas passaram a dividir os espaços, começando a ficar gradativamente mais perigoso, influenciando no acréscimo de acidentes.

Defronte a essa circunstância precisava-se impor e estabelecer normas no trânsito, por isso em 21 de setembro de 1966 institui-se o segundo Código Nacional de Trânsito pelo Decreto-lei 5.108, o qual permaneceu vigente durante 31 anos até que em 1997 foi promulgado o Código de Trânsito Brasileiro, o qual entrou em vigor no ano de 1998, vigorando até os dias atuais.

De acordo Brasil (2009) o Código de Trânsito Brasileiro de 1997, *"Considera-se trânsito a utilização das vias por pessoas, veículos e animais, isolados ou em grupos, conduzidos ou não, para fins de circulação, parada, estacionamento e operação de carga ou descarga"*. O acidente de trânsito é uma evento que atinge diretamente a pessoa, imputando fatores vinculados à morte, incapacidade física, danos materiais, sendo possível provocar sérios problemas psicológicos e muitas vezes de difícil superação. Os acidentes são fenômenos ocasionais, imprevistos e de vários fatores, ou seja, eles não ocorrem assiduamente e propendem a ser inesperados em relação ao local e hora, sendo que cada acidente pode ser considerado uma cadeia sucessiva de fatos.

2.2 Visão geral da análise multivariada

Com o avanço da tecnologia computacional, hoje em dia é possível analisar grande massa de dados, principalmente com grandes quantidades de variáveis. Esse impacto, possibilitou aos pesquisadores das diversas áreas do conhecimento científico obter conclusões mais precisas por meio dos seus modelos teóricos (HAIR; ANDERSON, 2005). Segundo autor supracitado, a análise multivariada tem a finalidade de avaliar ou examinar várias variáveis relacionadas simultaneamente, sendo todas consideradas importantes a princípio. Todas as variáveis devem ser aleatórias e estabelecendo uma relação mutua entre si, de forma que seus resultados não podem ser explicados separadamente. Desse modo, seu objetivo é reduzir os dados ou simplifica-los estruturalmente, ordenar e agrupar, investigar a dependência entre variáveis, predizer e construir testes de hipóteses.

Segundo JOHNSON e WICHERN (1998), se um observador quiser compreender

algum acontecimento ou evento, basta colher uma amostra com n indivíduos ou a unidade que queira observar e registrar suas medidas ou valores de um número $p \geq 1$ de variáveis de interesse. Denomina-se dados multivariados ao grupo de medidas ou valores observados das p variáveis nos n indivíduos ou unidade de medida. A Tabela 1 apresenta os n indivíduos ou itens com várias váriaveis. A notação x_{ij} é usada para indicar um valor particular da k-ésima variável mensurada na j-ésima unidade amostral ou de medida.

Tabela 1 – Organização das n medidas nas p variáveis.

Unidade de Observação	Variáveis ou Características					
Indivíduo ou Ítem	Var_1	Var_2	$\cdots$	Var_k	$\cdots$	Var_p
1	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
j	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Como exemplo de dados multivariados tem-se o trabalho de Bumpus (1898), que obteve os dados de 49 pássaros, em que estudou as medidas do corpo que foram o comprimento total, a extensão solar, o comprimento do bico e cabeça, comprimento do úmero e o comprimento da quilha do esterno. Desta forma os indivíduos foram considerados como sendo cada pássaro ou pardocas e as variáveis as medidas do corpo das pardocas (MANLY, 2008).

Outro exemplo podemos encontrar no estudo de Hongyu, Sandanielo e Oliveira (2015), cujas variáveis foram obtidas de um banco de dados que foi retirado do SAS (2008), onde foi fornecido informações sobre a taxa de criminalidade de varias cidade dos Estados Unidos, as cidades foram New York, Los Angeles, Detroit, Washington, Hartford, Honolulu, Boston, Tueson, Portland, Denver, Chicago, Atlanta, Houston, Dallas, New Orleans e Kansas City, na qual foram estudadas: assassinato, estupro, roubo, assalto, arrombamento, pequeno furtos e roubo de veículos.

Na análise multivariada, quando se tem em mãos um conjunto de dados com várias variáveis é de suma importância que o mesmo apresente correlação entre elas, assim podendo aferir, explicar e predizer o nível de relacionamento entre variáveis estatisticamente. Uma vez verificado a existência de correlação entre as variáveis é preciso saber qual método será aplicada aos dados. A análise multivariada compreende dois métodos: Avaliação da Interdependência é aquela nenhuma variável ou grupo de variáveis é definida(o) como independente ou dependente, ao invés, o procedimento envolve a análise simultânea de todas as variáveis dos dados, como por exemplo, na análise de componentes principais. O outro método é a avaliação de dependência que estuda a correlação de uma ou mais variáveis associadas às outras, como pode ser feita por meio de uma análise de regressão.

Outro ponto é identificar se as variáveis são métricas ou não métricas. As variáveis métricas são as que podem quantificar, medir que são utilizados os dados quantitativos as não métricas são as variáveis que são atribuídas ausência ou presença de alguma característica no caso os dados qualitativos ou categóricas (HAIR; ANDERSON, 2005).

Existem diversos métodos multivariados, os quais são utilizados conforme o objetivo de estudo. De acordo com Mingoti (2005), a metodologia multivariada é dividida em dois grupos. O primeiro grupo consiste em técnicas para simplificar a estrutura de variabilidades dos dados, para estudar a relação das variáveis entre se dentro do mesmo conjunto ou entre indivíduos. Este primeiro grupo engloba os seguintes métodos:

i) **Análise de componentes principais e análise dos fatores**: o objetivo principal é explicar a estrutura de variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, por meio da construção de combinações lineares das variáveis originais. Interessa-se obter redução do número de variáveis a serem avaliadas e interpretadas as combinações lineares construídas. Já a fatorial resulta em uma estrutura que tem como finalidade originar escalas variadas.

ii) **Analise de correlação canônica**: objetivo é relacionar concomitantemente várias variáveis métricas sendo dependentes e independentes.

iii) **Analise de agrupamento**: seu objetivo é classificar uma amostra de indivíduos ou objetos ou características em um pequeno número de grupos que não podem ocorrer ao mesmo tempo, de acordo com a semelhança entre os indivíduos.

iv) **Análise descriminante**: seu objetivo é assimilar diferenças entres grupos ou conjuntos e prever a probabilidade de que um indivíduo ou objeto corresponderá a uma classe ou grupo particular de fundamentos de acordo com as variáveis independentes métricas.

v) **Análise de correspondência**: uma técnica de correlação recentemente criada para facilitar tanto na redução da dimensão da classificação de objetos (exemplo: produtos, pessoas) em um conjunto de características quanto ao mapeamento perceptual (gráficos) de objetos relativos a essas características.

No segundo grupo, fazem parte os métodos de estimação de parâmetros como:

i) **Teste de hipótese**: Ao encontrar o modelo que consiga levantar hipóteses em função dos parâmetros estimáveis.

ii) **Análise de variância**: Permite estimar e comparar médias de duas ou mais populações normais multivariadas independentes (HAIR, 2005).

iii) **Covariância**: Pode ser utilizada em uma junção da MANOVA para remover o efeito de quaisquer variáveis dependentes. Serve para aferir o nível de associação linear entre duas variáveis aleatórias.

iv) **Regressão multivariada**: Seu objetivo é prever as modificações nas variáveis dependentes como respostas a alterações nas variáveis independentes.

Dentre os métodos citados acima, este trabalho so irá utilizar a análise de Componentes Principais, no qual iremos demonstrar a construção e determinação dos componentes, assim como a aplicação nos dados obtidos de acidentes de trânsito no município de Campina Grande-PB.

2.2.1 Análise de Componentes Principais

Em 1901 Karl Pearson, com um pensamento futurista e com a finalidade de resolver problemas naquela época, desenvolveu um método matemático que tinha por finalidade estudar um banco de dados com duas ou até três variáveis. Mas adiante em 1933, Hotelling continuou o estudo inciado por Karl Pearson, dando origem a um dos métodos considerados mais simples da técnica multivariada, a análise dos componentes principais (*Análise de Componentes Principais - ACP*). Porém não se tornava muito simples as análises, pois a tecnologia também estava em fase de desenvolvimento, e os cálculos eram enormes e feitos manuais. Segundo Mingoti (2005), a análise de componentes principais tem por finalidade construir combinações lineares das variáveis originais, formadas de p-variáveis aleatórias e explicar a estrutura de variância e covariância de um vetor aleatório. Estas combinações lineares não correlacionadas são chamadas componentes principais.

Além disso os componentes principais tem sua importância avaliada pela sua contribuição. Em síntese, tal importância é mensurada pela proporcionalidade da variância total apresentada pelo componente. A soma dos autovalores k representa a proporção de informação contida na redução de p para k dimensões. Através disso pode-se decidir quantos componentes utilizar na análise (VARELLA, 2008).

Com isso, a análise de componentes principais busca reduzir um grande número de variáveis originais há um pequeno número de variáveis transformadas. Assim, quando as variáveis apresentarem uma correlação alta, possivelmente essas variáveis serão representadas por dois ou três componentes principais por meio da diagonalização de matrizes simétricas semipositivas definidas. Desse modo podemos calcular simplesmente os componentes principais, e utiliza-los em diversas áreas científicas possibilitando as pesquisadores realização de inferências em relação a algum evento ou acontecimento (MANLY, 2008).

Assim como todo método estatístico, na análise de componentes principais se faz necessário apresentar alguns pontos que sejam, positivos ou negativos, os quais são:

i) Não exige que os dados tenha distribuição normal. Porém se os componentes forem de populações normais multivariados são apresentados pela elipsóide de densidades constantes (MINGOTI, 2005).

ii) Cada componente principal não se associa aos outros componentes, ou seja, não apresenta nenhuma dependência ou relação de causa-efeito entre as variáveis. Desta forma, não há modelo de medida casual (LATTIN; CARROLL; GREEN, 2011).

iii) Para Ferreira (2011) a análise de componentes principais deve ser vista como uma técnica exploratória (intermediaria) utilizada para facilitar enormes investigações científicas. Para JOHNSON e WICHERN (1998) *"A análise de componentes principais é um meio para se chegar a um fim e não um fim em si mesmo"*.

iv) A análise de componentes principais necessita apenas da matriz de covariâncias $\sum$ ou da matriz de correlação ρ (JOHNSON e WICHERN, 1999).

2.2.2 Construção dos Componentes Principais

Considere $\mathbf{X} = (X_1, X_2, ..., X_p)^T$, um vetor aleatório composto de variáveis de interesse com um vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^T$ e matriz de covariância $\boldsymbol{\Sigma}_{p\times p}$. Então sejam $\lambda_1 \geq \lambda 2 \geq ... \geq \lambda_p$ autovalores da matriz de covariância $\boldsymbol{\Sigma}_{p\times p}$ e seus autovetores sendo apresentado por $e_1, e_2, ..., e_p$, isto é os autovetores e_i estabelecem a í-ésima combinação linear para $i = 1, 2, ..., p$ (MINGOTI, 2005). Assim o i-ésimo componente principal $\mathbf{Y}_i$ é representado por

$$\mathbf{Y}_i = \mathbf{e}_i^T\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p.$$

Basicamente a definição dos componentes principais é lançar os pontos coordenados originais em um plano, superestimando a distância entre os mesmos, ou seja, maximizando a variabilidade do i-ésimo componente principal $\mathbf{Y}_i$. A variância e covariância do componente principal $\mathbf{Y}_i$, são respectivamente:

$$Var(\mathbf{Y}_i) = Var(\mathbf{e}_i^T\mathbf{X}) = \mathbf{e}_i^T Var(\mathbf{X})\mathbf{e}_i = \mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_i,$$

covariância entre $\mathbf{Y}_k (i \neq k)$ e o componente principal $\mathbf{Y}_i$ que será

$$Cov = (\mathbf{Y}_i, \mathbf{Y}_k) = Cov(\mathbf{e}_i^T\mathbf{X}, \mathbf{e}_k^T\mathbf{X}) = \mathbf{e}_i^T Var(\mathbf{X})\mathbf{e}_k = \mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_k.$$

Então, devemos elevar ao máximo a variância $Var(Y_i) = \mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_i$ com relação ao vetor $\mathbf{e}_i$, com excepção de $\mathbf{e}_i^T\mathbf{e}_i = 1$. A medida que $\mathbf{e}_i$ aumenta, a variância do componente principal também crescerá para o infinito. Segundo JOHNSON e WICHERN (1998) o primeiro componente principal será uma combinação linear $\mathbf{e}_1^T\mathbf{X}$ que maximiza a variância

de $\mathbf{e}_1^T\ \mathbf{X}$ se $\mathbf{e}_1^T\mathbf{e}_1$ for $= 1$. Dessa forma utilizaremos a técnica de multiplicadores de Lagrange onde λ_i é o multiplicador para maximizar o vetor $\mathbf{e}_i$, de acordo com Ferreira (2011).

$$\max_{\mathbf{e}_i}[\mathbf{e}_i^T\mathbf{\Sigma}\mathbf{e}_i - \lambda_i(\mathbf{e}_i^T\mathbf{e}_i - 1)].$$

Dividindo a função original por $\mathbf{e}_i^T\mathbf{e}_i$ e ainda derivamos essa função e igualamos a zero, depois de algumas simplificações pode-se dizer que

$$Var(\mathbf{Y}_i) = \lambda_i \quad \text{e} \quad Cov(\mathbf{Y}_i, \mathbf{Y}_K) = 0, \quad i \neq k.$$

De acordo com (FERREIRA, 2011) a definição dos componentes principais é a obtenção dos autovetores $\mathbf{e}_i$ que são a rotação dos eixos coordenados das variáveis originais e os autovalores λ_i que são os novos eixos coordenados, com $\mathbf{e}_i$ e λ_i variando $(1, 2, ..., p)$. Os componentes por serem ortogonais representam uma rotação mais rígida, de forma que escolhemos o componente principal de maior variabilidade ($\max_i \lambda_i$) e assim sucessivamente até o componentes de menor variabilidade, se ordenarmos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, sejam os componentes principais $Y_1 = \mathbf{e}_1^T\mathbf{X}, Y_2 = \mathbf{e}_2^T\mathbf{X}, ..., Y_p = e_p^T\mathbf{X}$, teremos tantas variáveis originais quanto componentes.

Pela decomposição espectral da matriz $\mathbf{\Sigma}$, em que $\mathbf{\Sigma} = \mathbf{P\Lambda P}^T$, que $\mathbf{P}$ é uma matriz ortogonal composta por autovetores e $\mathbf{\Lambda}$ a matriz diagonal de autovalores de $\mathbf{\Sigma}$ (JOHNSON; WICHERN, 1998). Assim,

$$tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P}^T) = tr(\mathbf{\Lambda}) = \sum_{i=1}^{p} \lambda_i.$$

Mas, $tr(\sum)$ é a soma dos elementos da diagonal,

$$tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda}) = \sum_{i=1}^{P} \sigma_{ii} = \sum_{i=1}^{P} \lambda_i,$$

Em que, a variância original populacional que são $\sigma_{11}+\sigma_{22}+...+\sigma_{pp}$ é a variabilidade total explicada pelos k-ésimo componente principal que é $= \lambda_1 + \lambda_2 + ... + \lambda_p$. Desta forma é representado por ,

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}, k = 1,2, \ ... \ , \ p.$$

Desta maneira os componentes podem ser representados por um, dois ou até mais componentes sem perda das informações. Segundo Ferreira (2011), se $k < p$ componentes principais para tentar gerar um modelo mais tranquilo para a matriz de covariância populacional, é preciso utilizarmos um critério para sabermos quanto da variabilidade foi explicada por ele, na forma de porcentagem. Para isto, considere o vetor $\mathbf{Y} = [Y_1, Y_2, ..., Y_p]^T$ que pode ter sua forma vetorial $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$, onde $\mathbf{P}$ é ortonormal ($\mathbf{P}^{-1} = \mathbf{P}^T$) e $\mathbf{X}$ pode ser obtida pela transformação não singular, representando então

$$\mathbf{X} = \mathbf{PY}.$$

Desta maneira, reduziremos os Componentes Principais (variáveis originais transformadas), para $k < p$, obtivermos o vetor $\mathbf{Y} = [Y_1, Y_2, ..., Y_k]^T (k \times 1)$, onde os k primeiros autovetores da matriz $\mathbf{P}$ que sejam aproveitados para preencher a matriz P_k $(p \times k)$, obtivemos

$$\mathbf{Y} = \mathbf{P}_k^T \mathbf{X}.$$

Por meio da matriz generalizada e utilizando a decomposição por valor singular, será fácil perceber que usa inversa é a própria matriz $\mathbf{P}_k^T = \mathbf{P}_k$. Assim pode-se obter as variáveis originais com determinado nível de precisão, o que propende da forma estrutural parcimoniosa ajustada da matriz de covariância, no qual predizemos as variáveis populacionais desta seguinte forma:

$$\tilde{\mathbf{X}} = \mathbf{P}_k \mathbf{Y},$$

e a $Cov(\mathbf{Y})$ é dada por

$$Cov(\mathbf{Y}) = Cov(\mathbf{P}^T \mathbf{X}) = \mathbf{\Lambda}$$

Como a variabilidade total é igual ao $tr(\mathbf{\Sigma})$, então, $Cov(\mathbf{Y}) = \mathbf{\Lambda}_k$, em que

$$\mathbf{Y} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

Por conseguinte, a variabilidade total do vetor $\mathbf{Y}(k \times 1)$ é o $tr(\mathbf{\Lambda}_k) = \sum_{i=1}^{k} \lambda_i$. A variabilidade total explicada pela proporção acumulada dos componentes principais em porcentagem, é descrita desta forma

$$\rho_k^2 = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \sigma_{ii}} \text{x } 100.$$

Para adquirir o k-ésimo é formalmente mostrada por

$$\mathbf{P}_k^2 = \frac{\lambda_k}{\sum_{i=1}^{p} \sigma_{ii}} \text{x } 100.$$

De acordo com Ferreira (2011), por meio da proporção acumulada da explicação da variância total, podemos determinar quantos componentes devemos obter ou representar. Há indícios científicos ou experiências vividas que para $k < p$ de componentes principais seria necessário utilizarmos componentes que expliquem pelo menos 70% da variabilidade dos dados. Como os componentes são ortogonais então $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ são reciprocamente ortogonais, pois a $Cov(\mathbf{Y}) = \mathbf{\Lambda}$ que é uma matriz diagonal, logo são não correlacionadas. Para JOHNSON e WICHERN (1998), os componentes podem explicar as p variáveis originais sem muita perda de informação se sua variabilidade total populacional for (80% a 90%) e puder ser representada a um, dois ou três primeiros componentes.

Segundo Regazzi (2000) e Varella (2008) é apropriado padronizar as variáveis $X_i(i = 1, 2, 3, 5, \cdots, p)$ quando se tem um banco de dados com variáveis de diferentes medidas. No qual deseja-se expressar os componentes em escalas padronizadas e centradas à zero. Para isto, faz-se necessário definir as variáveis como unidades de desvio padrão, assim, é só diminuir a média e dividir pelo desvio padrão da forma que possuam média zero e variância 1.

$$\mathbf{Y}^* = \left[\frac{\sum_{i=1}^{p} e_{1i}(X_i - \mu_i)}{\sqrt{\lambda_1}} \frac{\sum_{i=1}^{p} e_{2i}(X_i - \mu_i)}{\sqrt{\lambda_2}} \cdots \frac{\sum_{i=1}^{p} e_{pi}(X_i - \mu_i)}{\sqrt{\lambda_p}} \right]$$

Em notação matricial, teremos

$$\mathbf{Y}^* = \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{P}^T (\mathbf{X} - \boldsymbol{\mu}).$$

Estes são os componentes principais padronizados, onde $\mathbf{Y}^*$ é a nova variável padronizada, $\mathbf{Y}_i$ é a variável resposta antiga. Sejam os componentes padronizados com $E(\mathbf{Y}^*) = 0$ e $Cov(\mathbf{Y}^*) = \mathbf{I}$, onde possa-se obter o nível de associação entre eles e as variáveis originais. Para isso é preciso obter as covariâncias entre os componentes pelas variáveis originais.

$$Cov(\mathbf{Y}, \mathbf{X}) = Cov(\mathbf{P}^T \mathbf{X},\ \mathbf{X}) = \mathbf{\Lambda} \mathbf{P}^T.$$

Assim as covariâncias entre o k-ésima variável populacional e o i-ésimo componente principal e dada pela multiplicação do i-ésimo autovalor de $\mathbf{\Sigma}$ e o k-ésimo autovetor, ou seja $Cov(Y_i, X_k) = \lambda_i e_{ik}$. Dessa forma podemos encontrar a matriz de correlações entre o vetor de variáveis originais e o vetor dos componentes principais.

$$\boldsymbol{\rho}_{Y,X} = \mathbf{\Lambda}^{\frac{-1}{2}} \mathbf{\Lambda} \boldsymbol{P}^T \boldsymbol{V}^{\frac{-1}{2}} = \mathbf{\Lambda}^{\frac{1}{2}} \boldsymbol{P}^T \boldsymbol{V}^{\frac{-1}{2}},$$

em que a diagonal $\sigma_{ii} = \mathbf{V}$. De acordo com JOHNSON e WICHERN (1998) forma escalar, da matriz de correlação da i-ésima linha e k-ésima coluna, é dada por

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, Y_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{iK}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}}$$

$$= \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \mathrm{i}, k = 1, 2, ..., p,$$

em que, $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \cdots, (\lambda_p, \boldsymbol{e}_p)$ são os pares de autovalores e autovetores da matriz de covariâncias $(\mathbf{\Sigma})$.

2.2.3 Determinação dos Componentes Principais

Quando o objetivo é reduzir a dimensão dos dados, isto é, sumarizar informações das p- variáveis em k componentes, para $k < p$, faz indispensável estabelecer critérios de seleção para o determinar os valore dos componentes principais. Para tal faz necessário

padronizar as variáveis, pois as diferenças de medidas de apenas uma parte das variáveis contidas no vetor original $\mathbf{X}$, causa mudanças nos componentes de $\mathbf{\Sigma}$ (FERREIRA, 2011).

Então os componentes principais de $\mathbf{\Sigma}$ não são invariantes as diferenças de escalas. Se aplicarmos uma transformação do tipo $\mathbf{X}^* = \mathbf{\Delta X}$, que $\mathbf{\Delta}$ representa uma matriz diagonal $p \times p$, notaremos que os componentes das correlações populacionais $\mathbf{\Sigma}$ e os da matriz de covariâncias do vetor $\mathbf{X}^*$, dado por $\mathbf{\Delta\Sigma\Delta}$, não são os mesmos, determinados pela equação característica da matriz de correlação $\mathbf{\Sigma}$, que será:

$$det[\mathbf{\Sigma} - \boldsymbol{\lambda}_i\mathbf{I}] = 0 \quad \text{e} \quad [\mathbf{\Delta\Sigma\Delta} - \boldsymbol{\lambda}_i\mathbf{I}] = 0,$$

Notemos pela equação que os componentes de $\mathbf{\Sigma}$ não possuem a mesma solução. Se optar pela matriz $\mathbf{\Delta}$ por $\mathbf{V}^{-\frac{1}{2}}$ e ainda desenvolver uma padronização de posição da seguinte forma matricial:

$$\mathbf{Z} = \mathbf{V}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$$

a matriz de covariância de $\mathbf{Z}$ será

$$Cov(\mathbf{Z}) = \mathbf{V}^{-\frac{1}{2}}Cov(\mathbf{X} - \boldsymbol{\mu})\mathbf{V}^{-\frac{1}{2}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{V}^{-\frac{1}{2}} = \boldsymbol{\rho},$$

em que,

$$\mathbf{V}^{-\frac{1}{2}}\text{será} = diag(\frac{1}{\sqrt{\sigma_{ii}}}).$$

Assim, teremos que a matriz de covariância do vetor aleatório padronizado $\mathbf{Z}$, será igual a matriz de correlação populacional $\boldsymbol{\rho}$ (JOHNSON; WICHERN, 1998). A matriz $\mathbf{P}_{p\times p}$ é a matriz de covariâncias de $\mathbf{Z}_i$. Seus autovalores são representados pela matriz $\mathbf{P}_{p\times p}$ serão $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ e os autovetores normalizados por $e_1, e_2, \cdots, e_p$, onde $\mathbf{e}_1 = [e_{i1}, e_{i2}, \cdots, e_{ip}]$ e $\mathbf{\Lambda}$ representando a diagonal de autovalores da matriz $\boldsymbol{\rho}$ (MINGOTI, 2005). Teremos o modelo dos componentes principais pelas variáveis padronizadas da seguinte maneira:

$$\mathbf{Y}_i = \mathbf{e}_i^T\mathbf{Z} = e_{i1}Z_1 + e_{i2}Z_2 + \cdots + e_{ip}Z_p,$$

A matriz de correlações populacionais será

$$\boldsymbol{\rho} = \mathbf{P\Lambda P}^T.$$

Segundo Ferreira (2011), as variáveis do vetor padronizado $\mathbf{Z}$ representam pelos componentes principais rotações rígidas dos eixos coordenados padronizados. O sentido de maior variabilidade é representado pelo primeiro eixo elevando ao máximo a distância entre os pontos, o segundo eixo é perpendicular ao primeiro, sendo o segundo maior de maior variabilidade remanescente no sistema coordenado, os outros eixos se definem de maneira similar. Assim os autovalores passam a ser:

$$\mathbf{\Sigma}_{j=1}^{p}\lambda_i = tr(\boldsymbol{\rho}) = p,$$

representa que a variabilidade total incluída nas variáveis padronizadas (p) é igual a variabilidade total contida nos componente principais. Temos que a covariância do vetor $\mathbf{Y}$ é dada por

$$Cov(\mathbf{Y}) = \mathbf{\Lambda},$$

onde, para o modelo reduzido $k < p$ a covariância vetor $\mathbf{Y}_{(k\times 1)}$ é

$$Cov(\mathbf{Y}) = \mathbf{\Lambda}_k.$$

Ou seja, o vetor $\mathbf{Y}_{(k\times 1)}$ é $tr(\mathbf{\Lambda}_k) = \sum_{i=1}^{k} \lambda_i$. Assim, a variação total das variáveis será explicada pelo modelo de k componentes principais, essa proporção acumulada de quanto que é explicado a variação total expressa em porcentagem é da seguinte maneira:

$$\rho_k^2 = \frac{\sum_{i=1}^{k} \lambda_i}{p} \times 100.$$

A explicação individual pelo k-ésimo Componente Principal será

$$P_k^2 = \frac{\lambda_k}{p} \times 100, \mathrm{k} = 1, 2, \cdots, p$$

em que, λ_k são autovalores de ρ.

Agora, para avaliarmos o grau de associação entre as variáveis originais padronizadas a covariância será

$$Cov(\mathbf{Y}, \mathbf{Z}) = \mathbf{\Lambda}\mathbf{P}^T.$$

Sendo que $Cov(Y_i, Z_k)$ é a covariância entre a k-ésima variável original padronizada entre o i-ésimo componente principal no qual seja $\lambda_i e_{ik}$ o produto entre o i-ésimo autovalor de ρ e o k-ésimo componente principal do i-ésimo autovetor. Então a matriz de correlação é dada por

$$\rho_{Y,Z} = \Lambda^{\frac{1}{2}}\mathbf{P}^T,$$

e o coeficiente de correlação dado da forma escalar por

$$\rho_{Y_i,Z_K} = \sqrt{\lambda_i e_{ik}},$$

função do fato de a variável padronizada Z_k possui variância unitária.

Se desejar obter uma redução de suas variáveis sem perda de informações a soma dos primeiros k autovalores representa a proporção de informação retida na redução de p para k dimensões (VARELLA, 2008). Assim podemos decidir quantos componente principais iremos usar na análise, isto é, quantos componentes principais serão utilizados para diferenciar os indivíduos. Para escolher ou determinar o número de componentes principais é preciso impor alguns critérios de escolha, os mais utilizados são:

i) **Gráfico dos autovalores ou Scree Plot (Gráfico de Cotovelo)**: No qual se plota no eixo das abscissas k componentes principais e no eixo ordenado o seu autovalor λ_k. Para JOHNSON e WICHERN (1998) o gráfico se deu ao parecer realmente um cotovelo, os de maior fator ou autovalores estão em linha reta decrescente e os de menor estão em uma linha paralela à abscissa, podendo decidir quais os componentes que irão ser descartados e quais os que irão permanecer. Na Figura 1 os dois primeiros componentes poderiam representar a variância total.

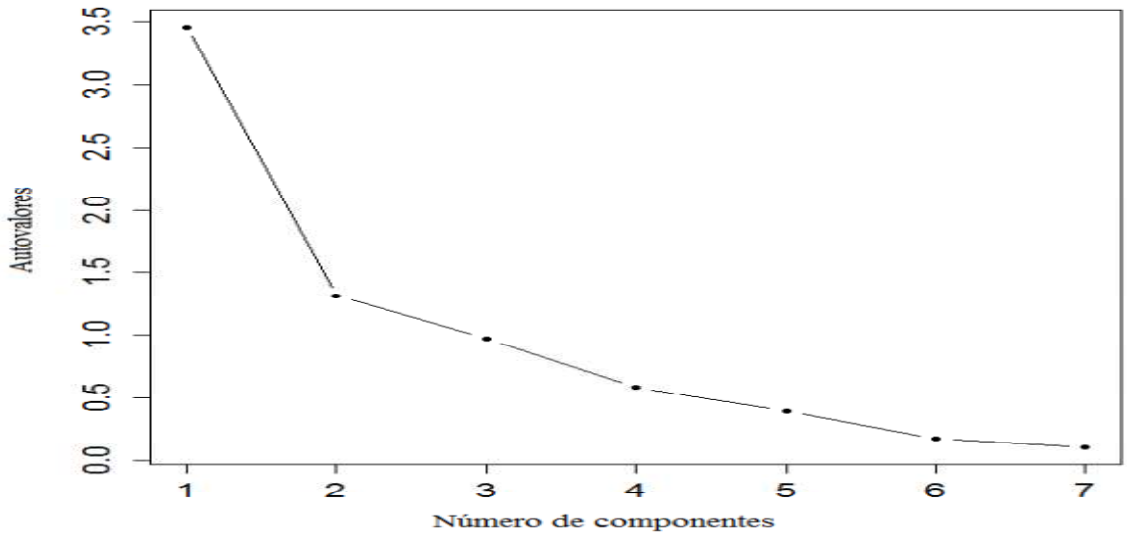

Figura 1 – ScreePlot ilustrativo de um exemplo com sete componentes principais

ii) **Critério de Kaiser (1958)**: no qual pretende preservar apenas os compoentes principais associados aos autovalores $\hat{\lambda}_i \geq 1$, que representem pelo menos a variabilidade das variáveis originais padronizadas (MINGOTI, 2005).

iii) **Teste de Esfericidade Bartlett (1950)**: Aplica o teste qui-quadrado de qualidade do ajuste e testa a hipótese nula no qual a matriz de correlação da população é igual a matriz identidade, no qual se aceitarmos a hipótese nula não seria correto a redução dos dados (LATTIN; CARROLL; GREEN, 2011).

Uma outra maneira de avaliar os resultados dos componentes principais é através do método biplot, que foi desenvolvido por Gabriel (1971), no qual representa graficamente os resultados ou a decomposição de valores singulares, em que o valor de cada elemento de uma tabela de dupla entrada pode ser visualizado pelo produto de vetores e pelo co-seno do ângulo entre dois vetores . Quando duas matrizes apresentarem o mesmo número de linhas e colunas, é possível multiplicá-las. A nova matriz gerada da multiplicação das duas anteriores assume o mesmo número de linhas e colunas de ambas as matrizes (YAN; KANG; MANJIT, 2003 apud HONGYU, 2015).

3 Material e métodos

O Sistema de Atendimento Móvel de Urgência (SAMU) vem realizando um importante trabalho na cidade de Campina Grande-PB e atende às vítimas de acidentes de trânsito auxiliando no atendimento e salvando vidas. Um estudo nesta cidade foi feito, com intuito de verificar e analisar os pontos de maiores índices de acidentes e explorar as possíveis causas, comparando-se por Zona da cidade, na intenção de servir para os órgãos responsáveis na adoção de medidas cabíveis para redução dos mesmos.

Para a realização deste trabalho foi utilizado dados obtidos do órgão do SAMU na cidade de Campina Grande - PB. Para a análise, utilizou-se apenas informações sobre acidentes de trânsito com vítimas. Realizou-se uma coleta de 730 boletins no SAMU, em que se considerou os acidentes de trânsito ocorridos só na Cidade de Campina Grande, Paraíba, no 1º semestre de 2014.

Em 2016, a cidade de Campina Grande tinha uma população estimada de 407.754 mil habitantes e possui uma extensão territorial no ano de 2015 com 593,023 km^2 de acordo com IBGE. Este município apresenta um total de 51 bairros oficiais e mais 3 distritos de acordo com a SEPLAN (Secretaria de Panejamento) de Campina Grande. A Figura 2 apresenta o mapa da cidade com suas divisões por Regiões. De acordo com a SEPLAN (Secretaria de Planejamento) da Prefeitura Municipal de Campina Grande, o município se divide em cinco Zonas (Central, Leste, Norte, Oeste e Sul).

Figura 2 – Regiões do Município de Campina Grande

Dos boletins foram retiradas as seguintes informações: Zonas (Central, Leste, Norte, Oeste e Sul) de acordo com a localização geográfica com base na SEPLAN do município de Campina Grande. Para identificar o perfil dos condutores foi coletado informações sobre:

- **Tipo Geral**: Compreende uma classificação sob uma perspectiva genérica, sem explicitar detalhes específicos do acidente apresentando várias categorias (Colisão, Atropelamentos, Queda, e outros).

- **Tipos de veículos**: As variáveis categorizadas foram (moto, carro, ..., carroça).

Com um intuito de se analisar algumas variáveis em função do tempo temos:

- **Mês**: Que foram os seis meses de pesquisa.
- **Dia do mês**: Os trinta dias do mês.
- **Dia da semana**: Tem sete níveis (segunda, terça, quarta, quinta, sexta, sábado e domingo).

Quanto aos Condutores coletamos informações das quais o condutor na hora do sinistro se comportava.

- **Usavam ou não equipamentos de segurança**.
- **Apresentava sinais de embriaguez**.

Para entendermos como foi o atendimento a essas vítimas dos acidentes foram coletadas as variáveis como:

- **Sexo**: Que foram masculino, feminino e não informado.
- **Condições das vítimas**: Se era passageiro, condutor, pedestre e não informado.
- **Atendimento às vítimas**: Como foi esse atendimento, se a vítima foi atendida no local significa que o acidente não foi tão grave, foi levada para o hospital e se houve óbito no local.

Como para o SAMU o importante é o socorro das vítimas, as fichas que apresentaram preenchimento incompleto, causando dificuldade na definição de alguns acidentes como por exemplo, acidente de moto, no qual não se especifica características detalhadas do acidente, restando-se a retirada dessa categoria como também as categorias acidente de carro e bicicleta, no qual essas categorias, não se dispõe de informações suficientes para estabelecer se o acidente envolveu outros veículos e como realmente o fez. Retiramos também os acidentes que ocorrem nos distritos e consideramos apenas os que ocorreram na Zona Urbana do município restando 597 observações.

A partir dos dados fornecidos pelo SAMU foi realizado inicialmente uma análise gráfica das informações, cuja finalidade foi observar o comportamento das variáveis disponíveis conforme o número de acidentes de trânsitos ocorridos. Foi utilizado o gráfico de barras para observar o perfil dos acidentes, verificar o número de acidentes de trânsitos ao longo do tempo (meses, dias do mês, dia da semana), estudar o comportamento dos

condutores e das vítimas. Em seguida foi realizado a análise de componentes principais com a finalidade de caracterizar e encontrar locais com maiores incidências de acidentes. Desta forma, a estrutura de dados multivariados utilizados na aplicação foi o conjunto de dados, cujas linhas da matriz foram as zonas e as colunas (variáveis) os tipos de acidentes (colisão, atropelamento, queda, capotamento e múltiplos).

A variável colisão é um embate entre dois ou mais corpos, por exemplo quando dois veículos se chocam. O atropelamento que vem da palavra atropelo, que pode ser uma colisão de um veículo com um pedestre ou um veículo com animal. Queda é um efeito de cair, que pode ser, cair da moto, caminhão e dentro outros. Capotamento a condição em que o veículo gire ou mude de posição sem colidir em nada e que fique com o teto encostado ao solo. A variável múltiplos pode ser vários acidentes que ocorreram de formas diferentes pode ser um tombamento que é o veículo virar para a direita ou esquerda, uma queda de carroça.

Os gráficos foram feitos utilizando a análise de componentes principais e o *biplot*, utilizando o pacote *stats*, *vegan* do software R versão 3.4.1.

4 Aplicações

A priore foi realizado uma análise descritiva do conjunto de informações, cuja finalidade é ter uma visão previa do comportamento destes dados.

4.1 Análise gráfica dos dados

A Figura 3, apresenta informações dos acidentes de trânsito dentro de cada Zona e observa-se que apesar da Zona central ser composta por somente três bairros que são Centro, São José e Prata, ainda conseguiu ultrapassar em percentual as Zonas Leste e Norte. A Zona que mais se destacou foi a Sul com 36,85% ocupando a primeira posição em relação aos acidentes. A Zona Oeste ficou em segundo lugar representando 26,47%, embora que a Zona Oeste apresente uma extensão territorial maior que as demais, não necessariamente ocorre-se maiores índices de acidentes, mesmo assim a Zona Sul conseguiu ultrapassar em número de acidentes.

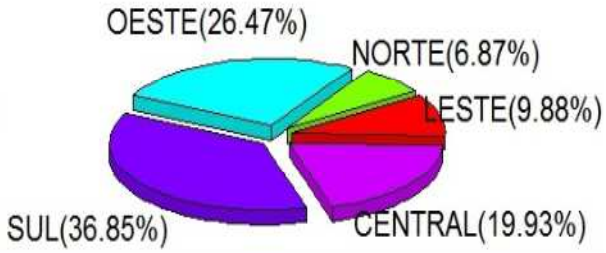

Figura 3 – Nº de Acidentes por Regiões

Na (Figura 4) as variáveis Colisão e a Queda foram os principais tipos de acidentes que mais ocorreram. A zona Sul e Oeste e Central se destacaram em todos os tipos de acidentes com maiores índices, diferentemente das Zonas Leste e Norte.

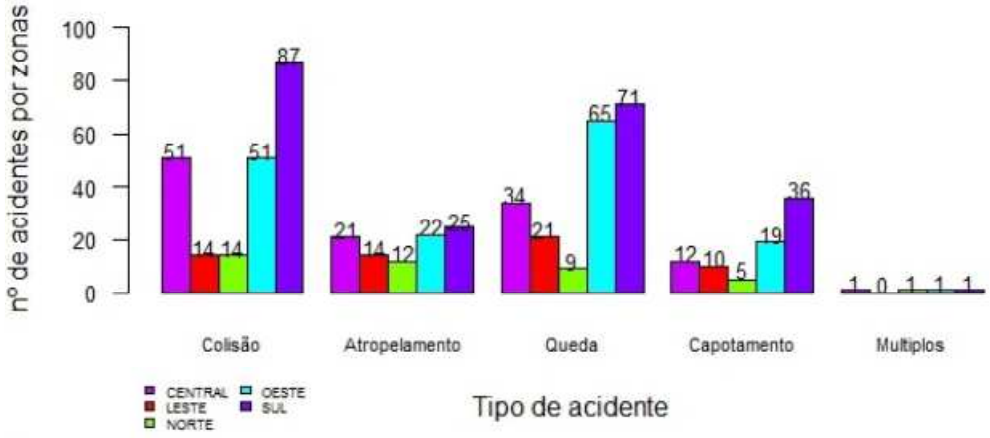

Figura 4 – Análise das Regiões por Tipo de Acidente

De acordo com a Figura 5, o tipo de veículo que mais se envolveu nos acidentes independente da Zona foi a motocicleta, representando na Zona Sul uma frequência de 136 motos. O segundo ficou automóvel com o maior índice na Zona Sul com uma frequência de 36. Percebeu-se também uma participação das bicicletas com uma frequência maior na Zona Oeste que foi 7 bicicletas envolvidas.

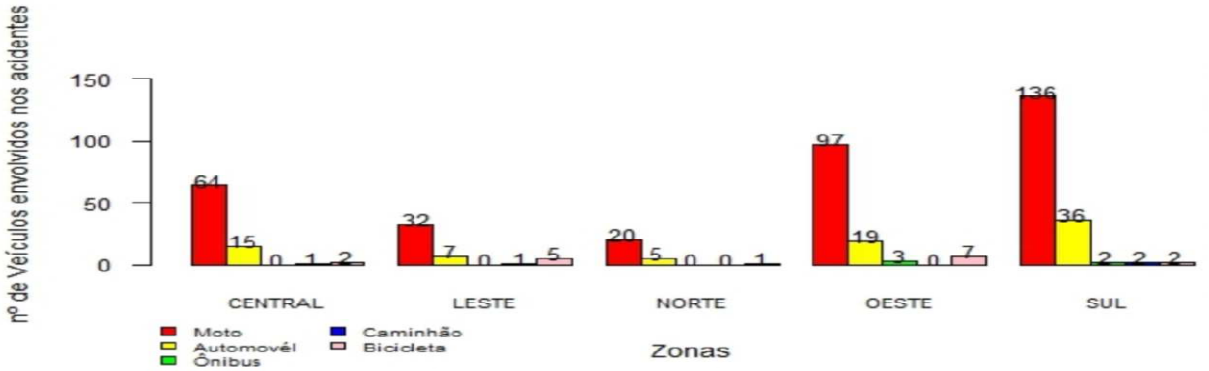

Figura 5 – Análise das regiões por tipo de veículo

A Figura 6 que mostra o primeiro semestre de 2014 em acidentes por Zona. Os meses que mais se destacaram em números de acidente foram Março e Abril, observou-se também que a Zona Sul se destacou em primeira posição quase em todos os meses, perdendo só no mês de Junho e possivelmente poderia ser justificado, por ser considerado o mês de festa. A Zona em que mais ocorreu acidentes com uma frequência de 36 foi a Zona Oeste com 6 acidentes a mais do que a Zona Sul.

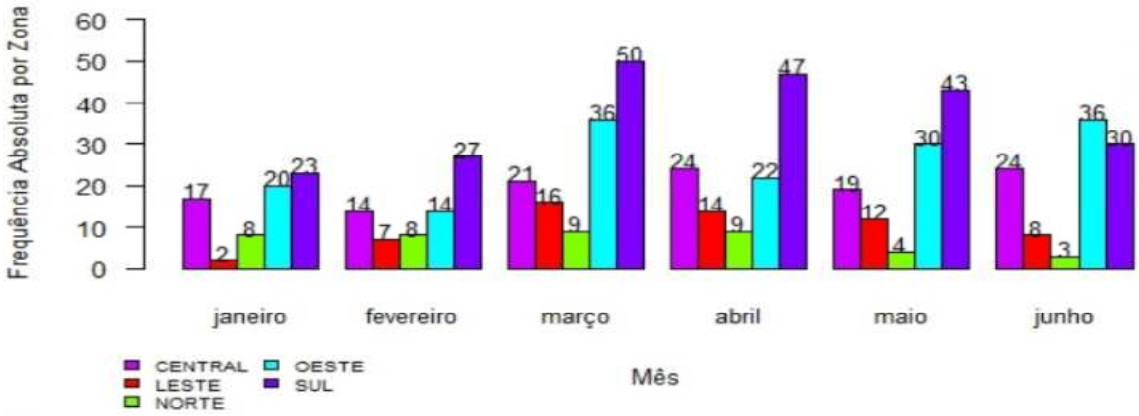

Figura 6 – Análise semestral por zona

A Figura 7 refere-se aos acidentes por zona e por dia do mês, de modo geral os acidentes apresentaram maiores incidências de acidentes nos dias 3, 5, 15 e 25. Ao observar cada zona, a Sul foi a que ocorreu um número maior de acidentes em relação as outras Zonas, e o dia em que ocorreu mais sinistros foi o dia 12. Já a Zona Oeste, o dia de maior incidência dos acidentes foi dia 25. Na Zona Central foi o dia 6 de maior número de

sinistros. NA Zona Norte o número de acidentes foi bem menor se comparado as outras Zonas, com poucas elevações e seu dia de maior incidência foi dia 11 e 14, já na Zona Leste o dia 16 foi o que mais ocorreu acidentes.

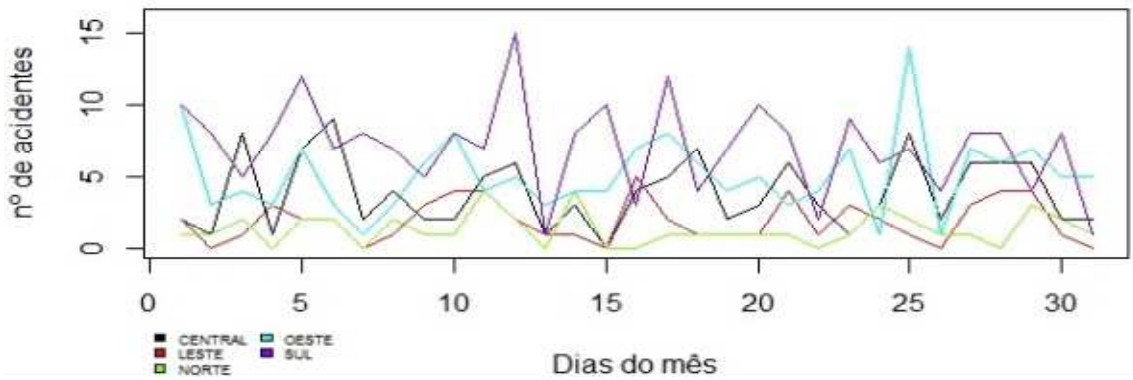

Figura 7 – Análise das Zonas por Dia do Mês

Ao analisar o dia da semana em que mais ocorre acidentes Figura 8, o domingo e a terça feira foram os dias em que mais se destacaram. O Domingo ficou em primeira posição nas Zonas Norte com uma frequência de (10) e Oeste (44), já a Terça feira foi o dia em que mais ocorreu acidentes na Zona Central (26) e Leste (15). Na Zona Sul o dia em que mais ocorreu esses acidentes foram no Sábado com uma frequência de 48 acidentes e o Domingo como sendo a segunda posição com 45 sinistros.

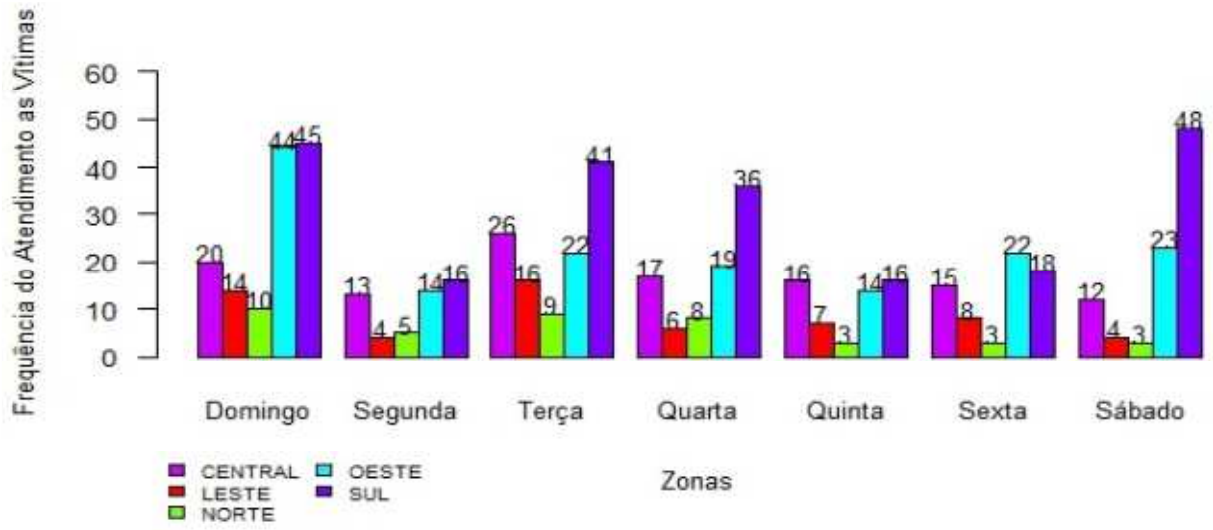

Figura 8 – Análise das zonas por dia da semana

Observa-se na Figura 9, os condutores envolvidos nos acidentes que informaram se utilizaram ou não equipamento de segurança e se apresentaram ou não sinais de embriagues. A grande maioria independente da Zona utilizaram equipamentos de segurança, como na Zona Sul em que 86 utilizaram na hora do acidente equipamentos de segurança. Porém ainda

existiu com pouca frequência, condutores que não utilizaram equipamento de Segurança prejudicando não só sua vida mais a vida de todos ao seu redor.

Ainda na Figura 9, o gráfico localizado no lado direito mostra, se os condutores apresentaram algum sinal de embriagues, nas Zonas Central e Oeste os condutores que apresentaram sinais de embriagues ultrapassaram os que não apresentaram nenhum sintoma. Na Zona central cinco condutores ingeriram algum tipo de bebida alcoólica e apenas 2 não apresentaram. Já na Zona Sul, dezesseis condutores não apresentaram sinais de embriagues para apenas 3 que apresentaram.

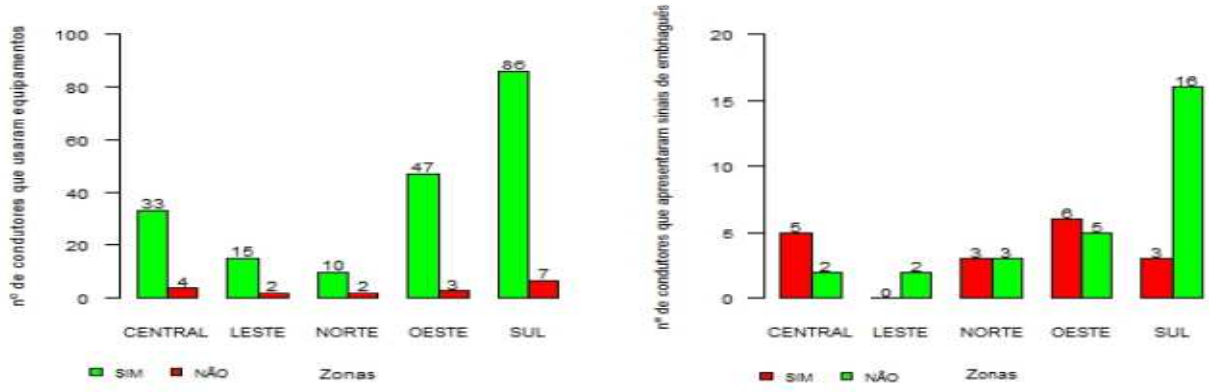

Figura 9 – Condutores que utilizaram ou não equipamentos de segurança e se apresentaram ou não sinais de embriagues por Zona

Estar embriagado e se envolver num acidente de trânsito é crime de trânsito, podendo o autor cumprir de seis meses a três anos de detenção de acordo com o Art. 306 do CTB(Código de Trânsito Brasileiro). Fica uma alerta para os condutores dirijam com atenção, preservem a sua vida. O comportamento das pessoas vitimadas nesses acidentes de trânsito. Como por exemplo,

Na Figura 10 em que, o sexo Masculino foi o que mais se feriu nos acidentes representando mais que o dobro em relação as mulheres, isso é independente de Zona, como podemos verificar na Zona Sul em que 157 homens ficaram feridos e apenas 47 mulheres foram vitimadas.

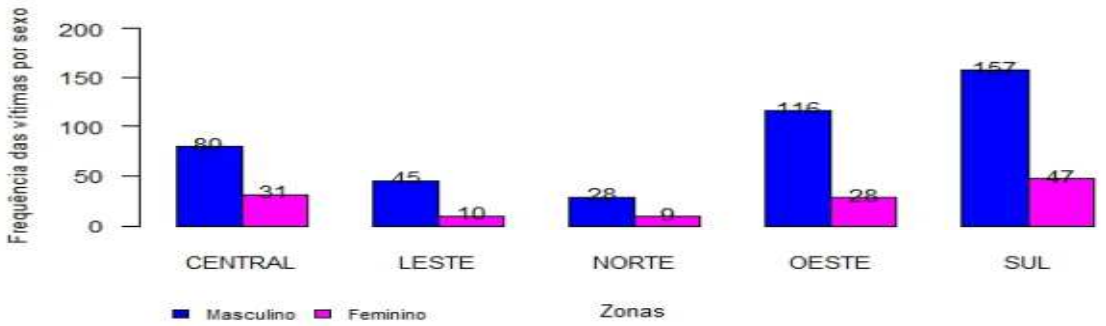

Figura 10 – Vítimas de acidentes por sexo e zona

Mostra-se na Figura 11 que dos acidentes de trânsito ocorridos, a maioria que se feriu foram os condutores dos veículos, no qual podemos verificar na Zona Sul e Oeste em que respectivamente tiveram uma frequência de 50 e 23 vítimas em que no momento dos acidentes eram condutores. Os pedestres deveram ter bastante atenção também, pois ficaram em segundo lugar nos acidentes que a vítima foi Pedestres. A Zona em que mais ocorreu acidentes com Pedestres foi a Zona Oeste com uma frequência absoluta de 11 pedestres feridos.

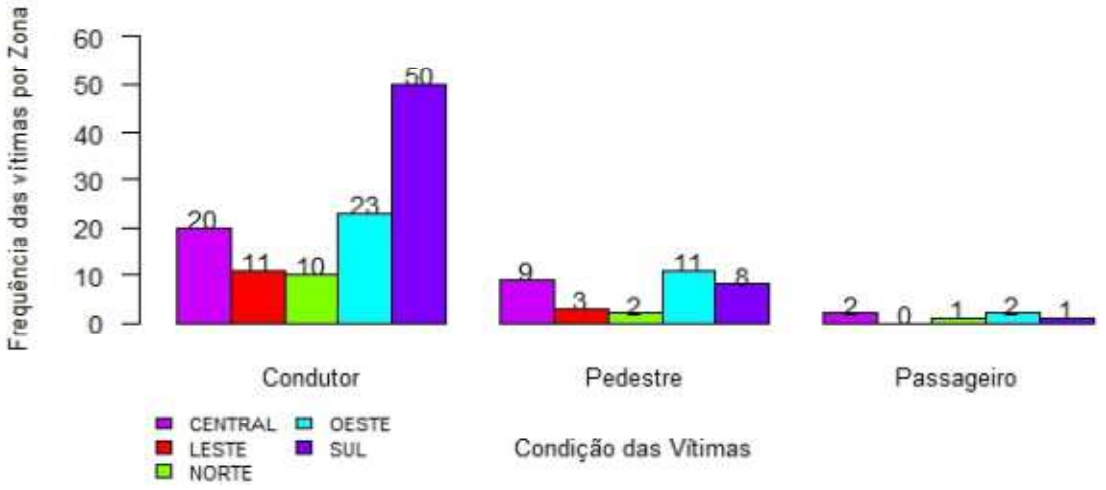

Figura 11 – Condições das vítimas por zona

Na Figura 13 se mostra como foi o atendimento a essa vítima de acidente de trânsito, se foi para o hospital, se ela foi a óbito no local, ou se o SAMU encaminhou para o hospital dentre outras. Observa-se que a grande maioria das vítimas nos acidentes foram encaminhadas para o hospital em todas as Zonas. Como no caso da na Zona Sul em que 189 pessoas feridas foram encaminhas para o hospital. Como ocorreu um índice muito

alto das pessoas que foram para o hospital, então retiramos esta variável para observar e analisar as demais.

De acordo com a Figura 13 se constatou que, na Zona Central e Oeste ambas obtiveram a mesma frequência de 6 acidentes nas variáveis Atendimento no Local e Removido por Terceiros. Na Zona Leste o Atendimento no Local à vítima foi o que mais ocorreu com uma frequência absoluta de 5 atendimentos. A Zona Norte o que mais se destacou foi à vítima Removida por Terceiros com 3 pessoas feridas. A Zona Sul o principal tipo de atendimento à vítima foi a Remoção por Terceiros com uma frequência de 8 pessoas removidas, também na Zona Sul houve um número maior de pessoas que se acidentaram e Evadiram do Local antes de chegar o SAMU que foram 4 pessoas feridas. De acordo com o SAMU as vítimas as vezes se evadem do local por medo de algo e acabam indo para suas residências e chamando o atendimento do SAMU.

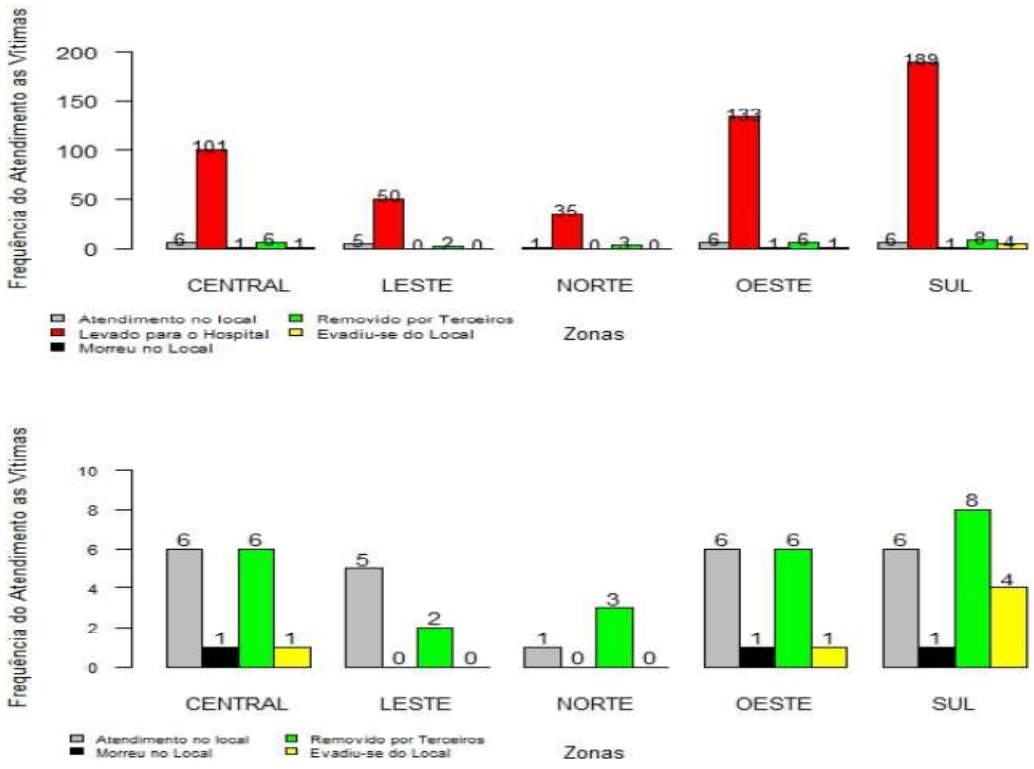

Figura 12 – Atendimento as vítimas de acidentes por zonas

Vale ressaltar que as Zonas que ocorreram morte por acidentes de trânsito, representam zonas mais perigosas em sinistros, valendo a atenção das pessoas e condutores que por ali transitam. Das Zonas consideradas mais perigosas em relação a ter ocorrido mortes foram a Zona Central, Oeste e Sul. A zona central foi a que mais se destacou, por haver um fluxo maior de veículos e pedestres transitando.

4.2 Aplicação da análise de Componentes Principais

Em uma análise multivariada é de suma importância verificar a correlação existente entre as variáveis em estudo, pois a técnica exige que as variáveis tenham uma correlação significativa. Geralmente, esta relação boa é representada por um valor da correlação acima de 0,5. Desta forma, foi utilizado o teste de correlação de Pearson nas variáveis duas a duas, e o resultado é apresentado na Figura 13. Observa-se que, a maioria das variáveis apresentaram uma correlação significativa com as demais variáveis, com exceção da variável múltiplos. Evidenciando que a variável múltiplos de nada influência na importância dos primeiros componentes principais.

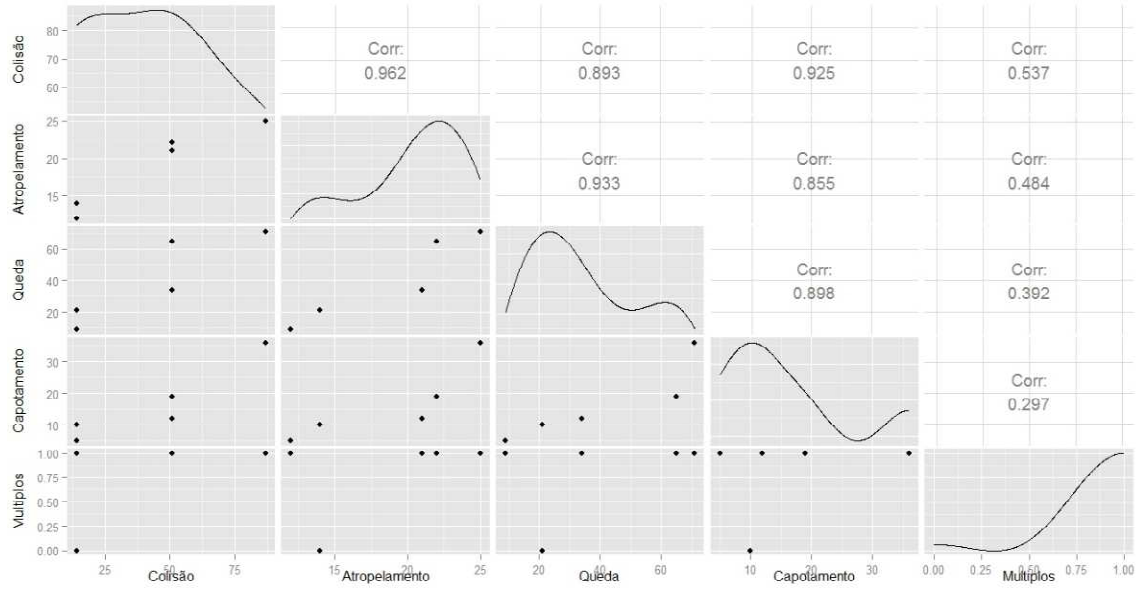

Figura 13 – Correlação entre os tipos de colisão.

Uma vez detectado correlação, pode-se utilizar o método multivariado de interesse. Nesta pesquisa, foi utilizado a análise de componentes principais e, conforme os resultados obtidos, na (Tabela 2) a partir dos autovalores, observa-se que, pelo critério de Kaiser (1958) um componente é suficiente para explicar a variabilidade total dos dados original padronizado, correspondendo a 79,60 % dessa variabilidade, pois apenas o primeiro componente apresentou valor maior que 1. Entretanto, observando a porcentagem de explicação acumulada percebe-se que dois componentes principais explicam 95,51% dessa variabilidade. Assim, o uso de dois componentes é recomendado para se obter uma conclusão mais acurada do estudo de pesquisa.

Sartorio (2008), em seu trabalho utilizou 13 variáveis referentes a percentagem das frações de animais e conforme os resultados concluiu-se que apenas três componentes principais eram suficientes para explicar 76,42% da variância total dos dados.

Tabela 2 – Componentes principais (CPs), autovalores (λ_i) e porcentagem da variância explicada e proporção acumulada (%) pelos componentes.

Componentes principais	Autovalores	Proporção	Proporção Acumulada (%)
PC1	3,98	79,60	79,60
PC2	0,79	15,91	95,51
PC3	0,13	2,69	98,20
PC4	0,09	1,80	100,00

Este mesmo resultado pode ser verificado por meio do gráfico *Scree Plot* (Figura 14), o qual apresenta os autovalores de maior importância em linha reta decrescente e os de menor importância começam a se posicionar em uma linha paralela à abscissa, ou seja, a partir do terceiro componente os autovalores se aproximam de zero passando a se estabilizarem, ficando bem próximos do eixo das abscissas. Desta forma, pode-se concluir que os dois primeiros componentes podem representar uma grande parte da variabilidade dos original padronizado.

Para FILHO et al. (2013), em seu estudo, verificou que por meio do *Scree Plot* apenas com um componente principal foi suficiente para explicar 90,57% da variabilidade das variáveis originais e a partir do segundo componente apresentaram valores abaixo de um, aproximando-se de zero.

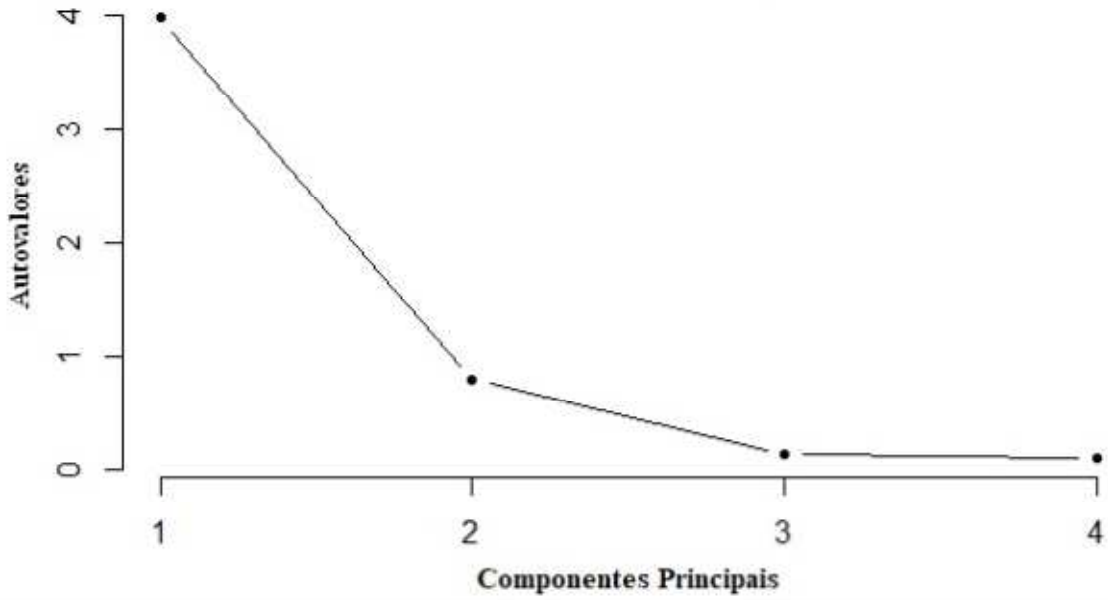

Figura 14 – Scree Plot de um conjunto de dados com 4 componentes principais

Portanto, para esta pesquisa a análise de componentes principais forneceu uma redução da dimensão das cinco variáveis originais para apenas dois componentes principais. Estes componentes são representadas a seguir.

$$CP1 = -0,494X_1 - 0,487X_2 - 0,476X_3 - 0,464X_4 - 0,277X_5 \quad (1)$$
$$CP2 = 0,004X_1 + 0,050X_2 + 0,186X_3 + 0,308X_4 - 0,931X_5 \quad (2)$$

Lembrando que X_1 é a variável colisão, X_2 é a variável atropelamento, X_3 queda, X_4 capotamento e X_5 múltiplos. Segundo Manly (2008) relata que, após a determinar os números de componentes principais a serem considerados na análise, é importante observar a magnitude das variáveis na seleção dos componentes, analisando o coeficiente de ponderação (variância dos componentes) em relação a variância das variáveis originais (correlação) de cada característica (Tabela 3).

Tabela 3 – Coeficientes de ponderação e correlação dos dois primeiros componentes principais.

Variável	Coeficientes de Ponderação		Correlação	
	PC1	PC2	PC1	PC2
colisão (X_1)	-0,494	0,004	-0,985	0,003
atropelamento (X_2)	-0,487	0,050	-0,971	0,045
queda (X_3)	-0,476	0,186	-0,951	0,166
capotamento (X_4)	-0,464	0,308	-0,926	0,275
múltiplos (X_5)	-0,277	-0,931	-0,553	-0,831

Uma das vantagens da análise de componentes principais é a representação gráfica conhecida como *biplot*, o qual tem a finalidade de mostrar de forma visual o comportamento das variáveis em estudo conforme o objetivo da pesquisa. Segundo Lattin, Carroll e Green (2011) cada ponto no gráfico representa um par de correlação, ou seja, cada variável principal correlacionada com um dos dois componentes principais indica quais delas estão mais associadas aos componentes.

Pode-se concluir que em ambas variáveis, a variabilidade é similar, isto pode ser observado pelo tamanho do vetor correspondente a cada uma delas. As variáveis X_1, X_2,X_3 e X_4 estão relacionadas com o primeiro componente (fato que foi observado na Tabela 3), pois seus vetores são refletidos em direção ao eixo do componente 1. Observa-se também a correlação existente entre as variáveis colisão, atropelamento e queda, pois formam ângulos agudos entres elas. Não existe correlação entre as variáveis capotamento e queda com múltiplos, pois forma um ângulo próximo de 90 graus.

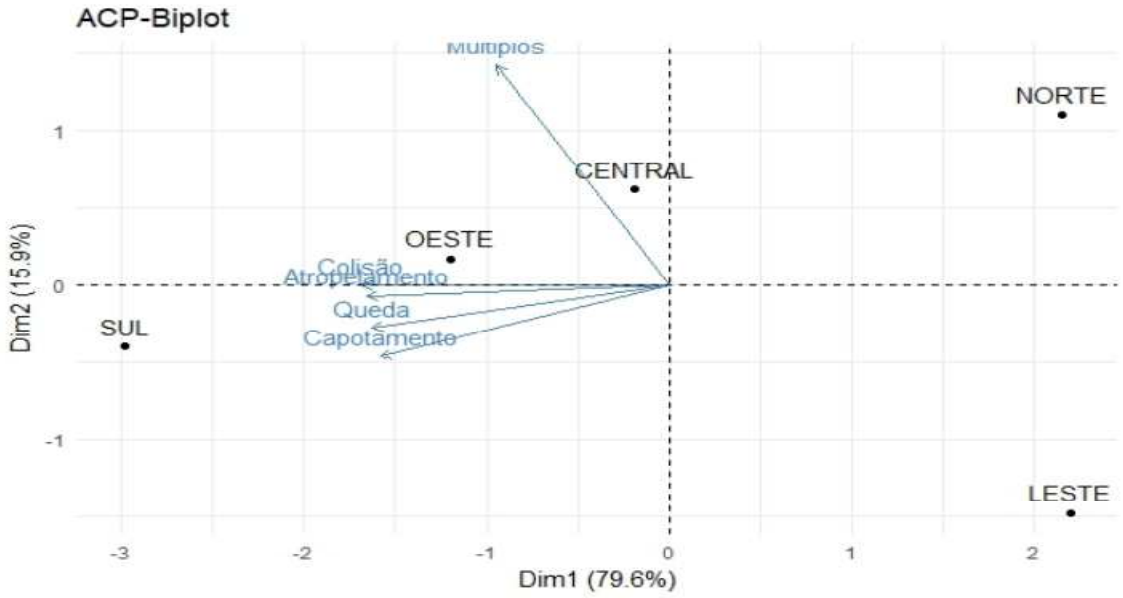

Figura 15 – *Biplot* $CP1 \times CP2$ de acidentes por zonas sobre o tipo dos sinistros

Por meio da Figura 15 pode-se concluir que, observando o componente 1 a Zona Central, Oeste e Sul apresentaram uma maior ocorrência de acidentes por colisão, atropelamento, queda e capotamento na cidade de Campina Grande. Entretanto, a Zona Norte e Leste não apresentam o mesmo comportamento. Pela componente 2, observa-se que a Zona Central e Norte apresentaram maior ocorrência de tipos de sinistros por capotamento e múltiplos.

As Zonas com maiores destaques que foram as Zonas, Central, Oeste e Sul, pode-se explicar essa quantidade de acidentes por nelas fazerem partes as avenidas de grandes extensões e que ocorrem bastante acidentes. A zona Central mesmo fazendo parte apenas com três bairros e sendo de pouca extensão, ainda conseguiu se destacar nos acidentes, diferentemente da Zona Sul e Oeste que abrange vários bairros e de terem uma maior extensão territorial.

5 Conclusão

Diante dos resultados obtidos, a análise gráfica mostrou que, as Zonas de maiores ocorrências de trânsito foram as Zonas Central, Oeste e Sul, O tipo de acidente em que mais ocorreu foi a colisão, queda, atropelamento, nos acidentes em todas as Zonas as motos estavam envolvidas, esses acidentes geralmente ocorrem nos finais de semana. Com base descritiva, a análise de componentes principais forneceu, de forma precisa, uma descrição dos acidentes de trânsito ocorridos em cada Zona da cidade de Campina Grande, considerando os tipos sinistros (colisão, atropelamento, queda, capotamento e múltiplos). Com apenas dois componentes principais foi possível explicar 95,51 % da variabilidade dos dados, considerando apenas as variáveis que apresentaram maior importância nos componentes sem perda de informação. A Zona Central, Oeste e Sul apresentaram uma maior ocorrência de acidentes por colisão, atropelamento, queda e capotamento na cidade de Campina Grande. A Zona Central e Norte apresentaram uma ocorrência maior de tipos de sinistros por capotamento e múltiplos. De forma geral, foi comprovado estatisticamente, tanto na análise descritiva como na análise de componentes principais, que os tipos acidentes que ocorreram com maior frequência na cidade de Campina Grande no ano de 2014 foram: colisão, atropelamento e queda e que as zonas de incidências foram Central, Oeste e Sul.

Referências

BRASIL, L. *Código de trânsito brasileiro*. [S.l.]: Senado Federal, Subsecretaria de Edições Técnicas, 2009. Citado na página 13.

DENATRAN. Relatório Estatístico, *Frota de veículos, por tipo e com placa, segundo as Grandes Regiões e Unidades da Federação - SET15*. 2015. Disponível em: <http://www.denatran.gov.br/estatistica/257-frota-2015>. Citado na página 11.

FERREIRA, D. F. *Estatística Multivariada*. 2. ed. [S.l.]: Editora UFLA, 2011. 676 p. Citado 4 vezes nas páginas 17, 18, 19 e 21.

FILHO, D. B. F. et al. Análise de componentes principais para construção de indicadores sociais. *Revista Brasileira de Biometria*, v. 31, n. 1, p. 61–78, 2013. Disponível em: <http://jaguar.fcav.unesp.br/RME/fasciculos/v31/v31_n1/A5_Dalson_Ranulfo.pdf>. Citado na página 34.

GAZIR, A. *Olavo Bilac era motorista no primeiro acidente do RJ*. 1998. Disponível em: <http://www1.folha.uol.com.br/fsp/especial/fj220116.htmhttp://www1.folha.uol.com.br/fsp/especial/fj220116.htm>. Citado na página 12.

HAIR, J.; ANDERSON, R. *TATHAN Ronald L. and William C. BLACK (2005). Análise Multivariada de Dados*. [S.l.]: Bookman, 2005. 593 p. Citado 2 vezes nas páginas 13 e 15.

HONGYU, K. *Comparação lavo do GGE biplot-ponderado e AMMI-ponderado com outros modelos de interação genotipo x ambiente*. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado na página 23.

HONGYU, K.; SANDANIELO, V.; OLIVEIRA, G. Análise de componentes principais: resumo teórico, aplicação e interpretação. 2015. Disponível em: <http://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398>. Citado 2 vezes nas páginas 11 e 14.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 6. ed. [S.l.]: Madison: Prentice Hall International, 1998. 816 p. Citado 8 vezes nas páginas 11, 13, 17, 18, 19, 20, 21 e 23.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Análise de Dados Multivariados*. [S.l.: s.n.], 2011. v. 475. Citado 3 vezes nas páginas 17, 23 e 35.

LESSA, D. Especial rodovias - as primeiras estradas brasileiras - (05' 49"). 2005. Citado na página 13.

MANLY, B. F. M. *MÉTODOS ESTATÍSTICOS MULTIVARIADOS: UMA INTRUDUÇÃO*. [S.l.]: Bookman, 2008. 229 p. Citado 3 vezes nas páginas 14, 16 e 35.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. [S.l.]: Editora UFMG, 2005. 297 p. Citado 5 vezes nas páginas 15, 16, 17, 21 e 23.

OLIVEIRA, F. *Dolo e Culpa Nos Delitos de Transito*. 1. ed. [S.l.: s.n.], 1997. ISBN 8524105224. Citado na página 12.

PR, D. *História do Trânsito*. 2006. Disponível em: <http://www.educacaotransito.pr.gov.br/modules/conteudo/conteudo.php?conteudo=141>. Citado na página 13.

PRF. Relatório, *Balanço de Atividades 2014*. 2015. Disponível em: <https://www.prf.gov.br/portal/sala-de-imprensa/releases-1/balanco-2014/view>. Citado na página 11.

SARTORIO, S. D. *Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R*. 130 p. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado na página 33.

VARELLA, C. A. A. Análise de componentes principais. *Seropédica: Universidade Federal Rural do Rio de Janeiro*, 2008. Disponível em: <http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf>. Citado 2 vezes nas páginas 16 e 20.

YAN, W.; KANG, M. S.; MANJIT, S. K. *GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists*. [S.l.], 2003. Citado na página 23.