Agrupamento de dados coletados sobre a rugosidade de uma amostra de calçadas na cidade de Joinville - SC - Brasil

Germano Augusto Metzner de Andrade
Universidade Federal de Santa Catarina
germano.andrade92@gmail.com

Pablo Andretta Jaskowiak
Universidade Federal de Santa Catarina
pablo.andretta@ufsc.br

Cassiano Augusto Isler
Universidade de São Paulo - Escola Politécnica (EPUSP)
cassiano.isler@ufsc.br

Andrea Holz Pfutzenreuter
Universidade Federal de Santa Catarina
andrea.hp@ufsc.br

AGRUPAMENTO DE DADOS COLETADOS SOBRE A RUGOSIDADE DE UMA AMOSTRA DE CALÇADAS NA CIDADE DE JOINVILLE – SC - BRASIL

G.A.M. Andrade, P.A.Jaskowiak, C.A. Isler e A.H. Pfützenreuter

RESUMO

A condição de pavimentação das calçadas é de interesse do público e das autoridades municipais para o processo de decisão e planejamento estratégico a fim de minimizar os custos de manutenção das vias. Este artigo apresenta resultados de um estudo de agrupamento de dados coletados sobre a rugosidade de uma amostra de calçadas da cidade de Joinville-SC-Brasil. Baseando-se em revisão bibliográfica referente a métodos de agrupamento, aplicou-se uma técnica de *k-means* com posterior validação por diferentes índices (PBM, Silhueta, Gamma, C-Index, Calinski Harabasz, Point Biserial e Xie Beni). A análise dos resultados provenientes dos índices de validação, sugerem que dois grupos são necessários para representar os padrões de rugosidade das amostras analisadas. Tal fenômeno pode ser explicado pela incapacidade dos atributos coletados em expressar com exatidão ou pelo detalhamento das características presentes nas calçadas analisadas.

1 INTRODUÇÃO

Caminhar é o método mais saudável e natural para o ser humano se locomover. Além de ser eficiente do ponto de vista energético, não prejudica o meio ambiente. Pelo fato de ser o modo de deslocamento de conexão entre modos de transportes o controle e manutenção das calçadas não deve ser excluído dos processos de planejamento urbano das cidades (BURDEN, 2001).

Atualmente as vias que deveriam facilitar a circulação dos pedestres e possibilitar a locomoção de pessoas com deficiência, na verdade tornam a experiência de seus usuários pouco prazerosa. A condição de pavimentação das calçadas é de interesse do público e das autoridades municipais. A informação sobre a qualidade da infraestrutura é necessária para o processo de decisão e, principalmente, para um planejamento estratégico a fim de minimizar os custos de manutenção das vias. Entretanto, o poder público não tem condições operacionais de verificar todas as calçadas que permeiam o meio urbano, de tal forma que se tem estudado métodos para medir a qualidade das calçadas e obstáculos existentes (ERIKSSON et al., 2008; GONZÁLEZ et al., 2008) .

Uma das maneiras possíveis de avaliar a condição das calçadas é observando as suas irregularidades, incluindo a forma da superfície, rachaduras, buracos e deterioração do pavimento. Um dos indicadores utilizados na literatura (ERIKSSON et al., 2008; GONZÁLEZ et al., 2008; MEDNIS, 2011) para medir a condição de qualidade das vias é a rugosidade, um parâmetro que não reflete apenas a qualidade do pavimento caminhável,

mas também indica a dificuldade de locomoção e influencia a segurança dos usuários em condições climáticas adversas.

O objetivo deste artigo é apresentar os resultados de um estudo de agrupamento de dados coletados sobre a rugosidade de uma amostra de calçadas na cidade de Joinville-SC-Brasil. Para atingir esse objetivo foram utilizados dados coletados por um acelerômetro acoplado a um smartphone, os quais foram convertidos em séries temporais e submetidas a uma técnica de agrupamento (*k-means*), com posterior validação por diferentes índices (PBM, Silhueta, Gamma, C Index, Calinski Harabasz, Point Biserial, Xie Beni).

2 REFERENCIAL TEÓRICO

Os smartphones são equipamentos promissores para coleta de dados devido ao seu baixo custo e a presença de diversos sensores acoplados como GPS, identificadores de luminosidade e acelerômetros.

Os arquivos de acelerômetros coletados para caracterização das calçadas são sequências de dados de uma variável em um intervalo de tempo. Essas sequências são denominadas séries temporais, um conjunto de pontos cuja ordem em relação ao tempo é relevante.

O processamento das séries temporais é feito a partir de algoritmos de agrupamento de dados, que segundo Facelli et al. (2005) são metodologias à análise exploratória de dados. Tais algoritmos fornecem um meio de explorar e verificar padrões presentes nos dados, organizando-os em grupos ou clusters.

As etapas que englobam o processo de agrupamento envolvem desde a preparação dos dados, execução do algoritmo de agrupamento propriamente dito até a interpretação dos grupos e a validação dos resultados. A relação entre as etapas do processo de agrupamento pode ser verificada Figura 1, explicadas separadamente após a ilustração conforme Jain et al. (1999), Faceli et al. (2005) e Barbara (2000).

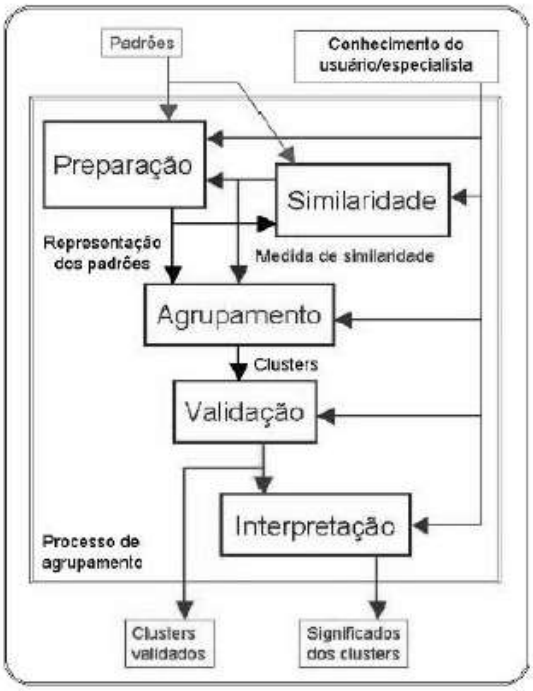

Figura 1 – Etapas do processo de agrupamento. Fonte: Faceli et al. (2005)

a) **Preparação dos dados:** Envolve a forma como os dados serão representados e a aplicação de transformações dos dados originais, como normalização e extração de características. Apesar de ser possível obter uma gama significativa de clusters válidos sem realizar esta etapa de preparação, na maioria dos casos os diferentes atributos que representam os padrões se apresentam em escalas diferentes. Quando os intervalos de valores dos atributos diferem muito, pode ser que um atributo domine o resultado do agrupamento fazendo com que ocorra uma tendência na relação entre os dados. Para solucionar esse problema, é comum a padronização de forma que os atributos estejam na mesma escala.

b) **Medidas de similaridade:** Consiste na definição apropriada ao domínio da aplicação. Em geral, uma medida de similaridade é fornecida por uma função de distância definida entre pares de padrões, sendo possível considerá-la sob aspectos conceituais (qualitativos) ou numéricos (quantitativos).

c) **Obtenção de agrupamentos:** Etapa de classificação dos dados para obtenção dos clusters. Cada algoritmo emprega um critério de agrupamento diferente (compactação, forma e equidistância são os mais comuns). Se, por acaso, os dados estão em equivalência com as exigências do critério estimado, então é possível afirmar que os clusters são válidos. Caso a resposta seja negativa é necessária uma interpretação dos resultados para identificação de possíveis inconsistências.

d) **Validação:** Etapa na qual a validade dos agrupamentos obtidos é questionada e verificada. A validação do resultado de um agrupamento, em geral, é feita com base em índices estatísticos, que indicam de uma maneira qualitativa, o mérito das estruturas encontradas. Cada índice avalia um parâmetro diferente do agrupamento tal que diferentes índices podem apontar características distintas para o mesmo grupo de dados.

Existem três tipos de critérios que podem ser aplicados: critérios relativos, internos e externos. Neste relatório não são aprofundados os conceitos de critérios internos e externos, pois estes não se aplicam ao problema abordado. Utiliza-se somente o conceito de critérios relativos, que consiste em validar quantos tipos de clusters há em uma determinada série temporal. Dependendo do método, o melhor número de clusters é dado pelo mínimo, máximo ou inflexão da curva observada.

e) Interpretação: Refere-se ao processo de examinar cada cluster com relação aos respectivos conteúdos para rotulá-los e descrever a sua natureza. Existem algoritmos capazes de realizar a interpretação para grandes quantidades de grupos, porém a análise de similaridade com os grupos básicos será analisada com base na porcentagem original de calçadas medidas.

3 METODOLOGIA

Para validação dos dados considerou-se que diferentes tipos de pavimentos podem causar interferências distintas no smartphone, dado que os materiais que compõem os pavimentos influenciam diretamente sobre a rugosidade da calçada. O método utilizado depende da produção dos resultados obtidos durante a medição. Nesta seção são descritos os experimentos realizados que dão suporte aos resultados obtidos.

3.1 Aquisição de dados

Para a aquisição de dados construiu-se um veículo que pudesse realizar medições nas calçadas. O modelo de veículo pensado foi um triciclo, pois o seu formato triangular evita que as rodas do veículo deixem de registrar imperfeições da via.

A realização das medições foi feita colocando-se o smartphone na parte posterior do veículo de testes (Figura 2). O smartphone utilizado para os experimentos foi um Motorola G4 Play (XT1600) e o aplicativo, disponível na comunidade do Google Play, utilizado para realizar a coleta das séries temporais foi o AndroSensor (FIV ASIM, 2015). A escolha do software deve-se aos recursos disponíveis, além de um controle de frequência de aquisição de dados ajustável.

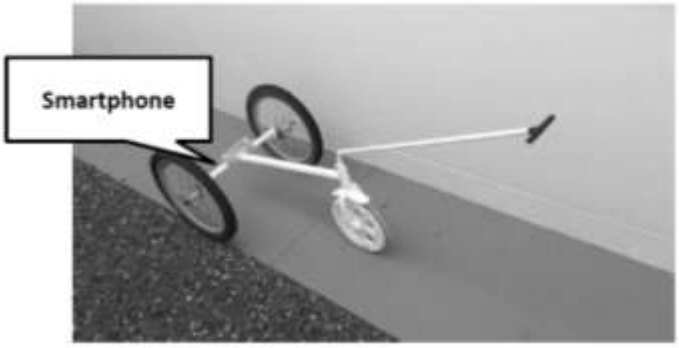

Figura 2 – Veículo construído para aquisição dos dados.

O sensor utilizado foi um acelerômetro embutido no smartphone considerando-se apenas variações identificadas no eixo Z. A orientação do smartphone é fixa de modo que a coordenada da aceleração e do veículo são as mesmas.

Na literatura (MOHAN et al., 2008; ERIKSSON, 2008; SPIRIC, 2014) não existe senso comum sobre uma taxa de aquisição de dados, sendo estabelecida a frequência de captura de dados em 200 Hz para quantificar os dados da rugosidade. Na perspectiva do usuário essa frequência é compatível com a distância de aproximadamente 1,5 metro. O veículo foi conduzido por dezoito tipos de pavimentos a fim de estabelecer uma base de dados para processamento, cujos materiais estão listados na Tabela 1.

Tabela 1 - Calçadas medidas e seus materiais.

MATERIAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CIMENTO	X							X				X						
CIMENTO E PEDREGULHO			X															
CIMENTO TEXTURIZADO									X									
PEDRA MIRACEMA		X			X					X								
PLACA TEXTURIZADA				X										X				
BLOCO INTERTRAVADO						X					X							
BLOCO HEXAGONAL													X		X			
BLOQUETE							X									X		
ASFALTO																	X	
PLACA DRENANTE																		X

Para evitar o viés do usuário que estivesse usando o veículo, estabeleceu-se que o mesmo percorresse o trajeto de ida e de volta na mesma calçada e com mesma velocidade, tendo-se um maior aproveitamento da via e consequentemente mais informações para análises. Algumas fotos das calçadas foram registradas para checagem e validação dos dados posteriormente, conforme exemplificado na Figura 3.

Figura 3 – Calçadas com bloco intertravado (esq.) e pedra miracema (dir.).

3.2 Pré-processamento de dados

Após coletar os dados em campo, os arquivos foram transferidos através da conexão Wi-Fi para computador sem nenhum processamento prévio. A leitura dos arquivos ".csv" gerados pelo celular ocorreu pelo programa estatístico R (THE R FOUNDATION, 2016) que uniu os arquivos existentes de cada calçada em uma única planilha. Dados de GPS e de identificação foram repartidos em vetores em relação à tabela original uma vez que, essas informações podem influenciar o algoritmo de agrupamento visto que estão correlacionados com os dados reais.

Com a intenção de captar as características da rugosidade, foi idealizada uma janela deslizante que percorria toda a série temporal com o tamanho aproximado de 200 amostras e defasagens de 100 medições entre elas. Isso significa que, em um exemplo prático, a primeira janela percorreria os dados entre os valores de 1 até 200 e a segunda janela entre os valores de 101 até 300.

As razões pelas quais se criou a janela com tais propriedades foram (i) verificar se anomalias como buracos e rachaduras ocorriam nessa relação entre a amplitude da janela e o intervalo de tempo correspondente; (ii) janelas de tempo relativas a intervalos de tempos maiores poderiam prejudicar a qualidade dos dados; e (iii) janelas de tempo com amplitudes menores comprometem o desempenho computacional durante as análises de agrupamento.

A Figura 4 ilustra uma série temporal obtida do acelerômetro pelo deslocamento do dispositivo de coleta de dados sobre uma calçada durante 3 segundos. Na sequência, a Tabela 2 apresenta a quantidade de janelas identificadas nas 18 calçadas percorridas ao

segmentar as respectivas séries temporais em janelas de tempo sucessivas de 200 observações, com defasagem de 100 observações e taxa de coleta de dados de 200 Hz.

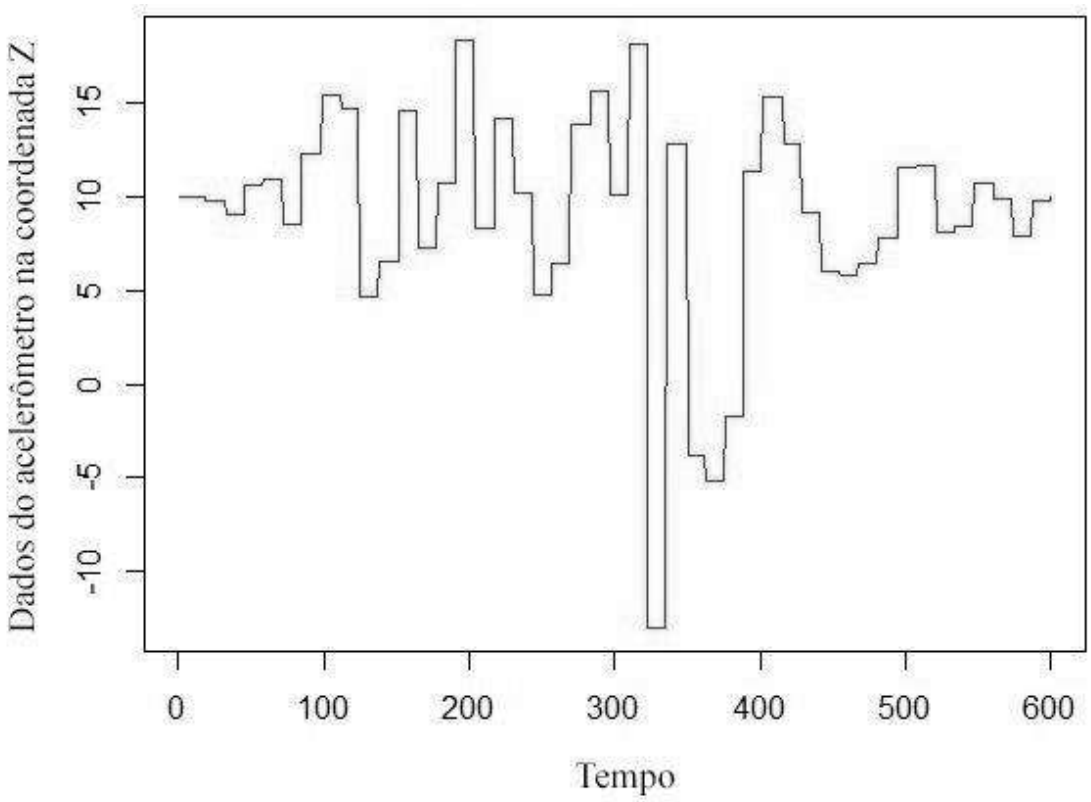

Figura 4 – Exemplo de dados extraídos com o smartphone em um intervalo de 3 segundos.

Tabela 2 - Quantidade de janelas extraídas de cada calçada.

Calçada	1	2	3	4	5	6	7	8	9
Janelas	150	150	154	146	138	147	194	163	150

Calçada	10	11	12	13	14	15	16	17	18
Janelas	129	111	162	119	193	186	118	141	162

Para avaliar o comportamento das séries temporais obtidas foram calculados 12 atributos para cada janela nas 18 diferentes calçadas: média; desvio padrão, mínimo; máximo; diferença absoluta; número de dados acima ou abaixo de 1 desvio padrão; número de dados acima ou abaixo de 2 desvios padrão; número de dados acima ou abaixo de 3 desvios padrão; soma da diferença absoluta de dois pontos consecutivos; desvio padrão da soma da diferença absoluta de dois pontos consecutivos; soma da diferença absoluta dos valores em relação à média; desvio padrão da soma da diferença absoluta dos valores em relação à média. Muitos desses atributos são utilizados na literatura e outros foram implementados para verificar se possuíam alguma relação com as séries temporais.

Após análise verificou-se pela matriz de correlação representada na Figura 5, que a relação entre alguns atributos era extremamente alta (maior que 0,9), indicando que tais características representavam as mesmas informações.

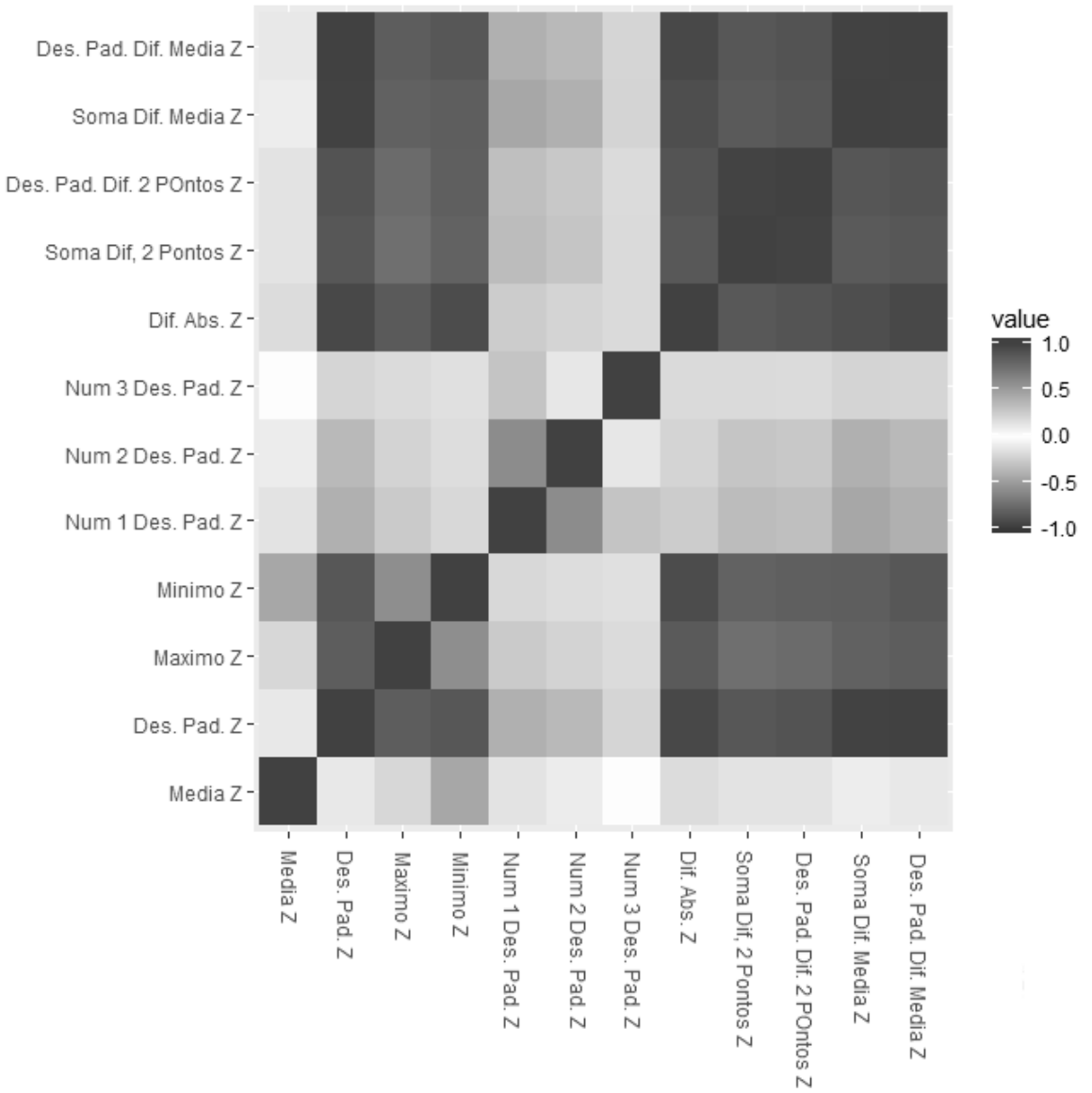

Figura 5 – Quadro de correlações de atributos

Para que o método de agrupamento não considerasse medidas correlacionadas foram removidos 4 atributos pertencentes às janelas deslizantes: média; diferença absoluta; soma da diferença absoluta dos valores em relação a média; e desvio padrão da soma da diferença absoluta dos valores em relação a média.

Na sequência, os valores dos atributos remanescentes (desvio padrão, mínimo, máximo, número de dados acima ou abaixo de 1 desvio padrão, número de dados acima ou abaixo de 2 desvios padrão, número de dados acima ou abaixo de 3 desvios padrão, soma da diferença absoluta de dois pontos consecutivos e desvio padrão da soma da diferença absoluta de dois pontos consecutivos) foram normalizados para que nenhuma das características dominasse a tendência dos dados antes da sequência de agrupamento.

Com os dados sintetizados por atributos normalizados aplicou-se a estratégia de agrupamento através do algoritmo k-means existente no programa estatístico R, escolhido por ser um dos algoritmos utilizados para análise de agrupamentos.

4 RESULTADOS E DISCUSSÕES

O número de grupos especificados pelos autores deste estudo foi de três clusters. Considerou-se que os dados obtidos poderiam ser representados por agrupamentos para condições “ruins”, “intermediárias” e “boas” de calçadas.

Com base apenas nas observações de agrupamento da Figura 6 nota-se uma relação direta entre os grupos 1 e 3. Com o crescimento do grupo 1 o grupo 3 também tende a crescer, mas, em grande parte, com uma porcentagem menor. A relação entre esses dois grupos e o grupo 2 é antagônica. Geralmente quando a porcentagem de dados deste grupo é alta os grupos 1 e 3 tendem a ser baixos, o inverso também se aplica.

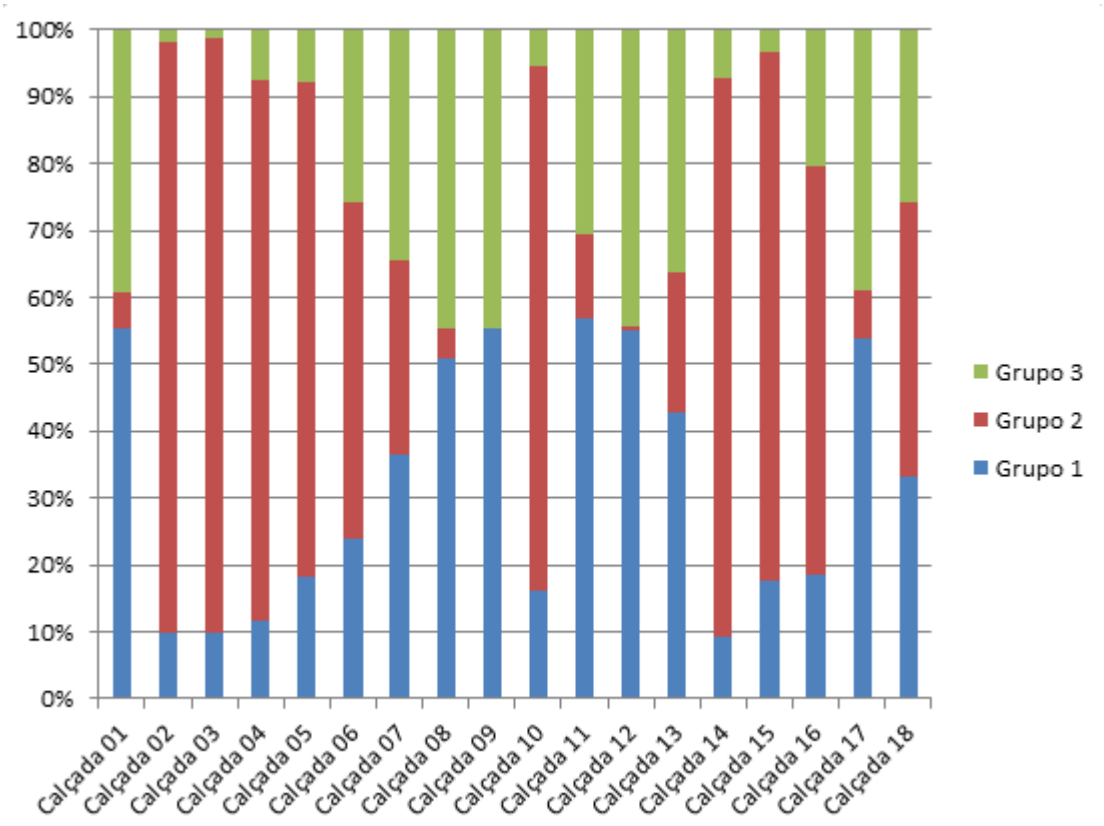

Figura 6 - Porcentagem de cada tipo de calçada nos seus possíveis clusters.

Para validar a quantidade de grupos estipulada computacionalmente foram usados os índices de validação "PBM", "Point Biserial", "C index", "Calinski Harabasz", "Gamma", "Silhueta" e "Xie_Beni" (Faceli et al., 2005), cujos resultados são caracterizados nos gráficos entre a Figura 7 e a Figura 13.

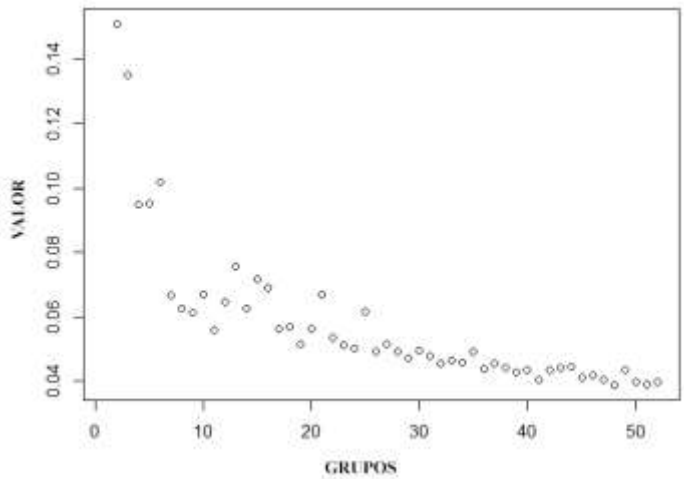

Figura 7 – Índice de validação (C INDEX)

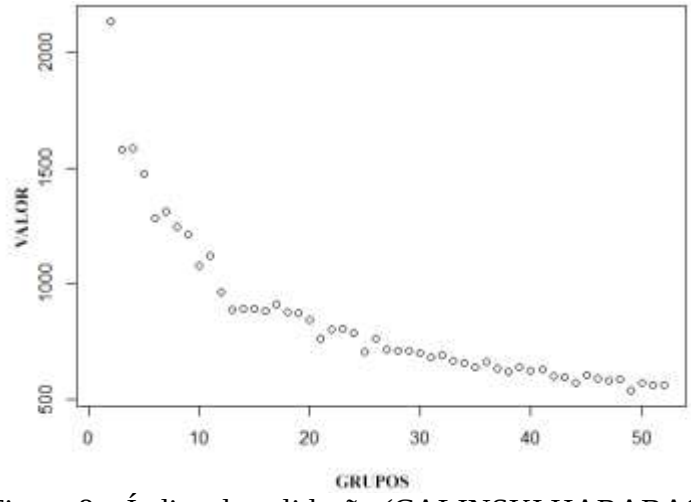

Figura 8 – Índice de validação (CALINSKI HARABASZ)

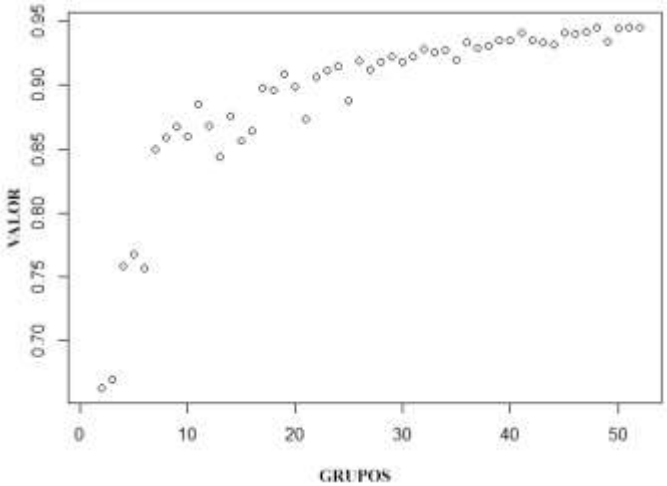

Figura 9 – Índice de validação (Gamma)

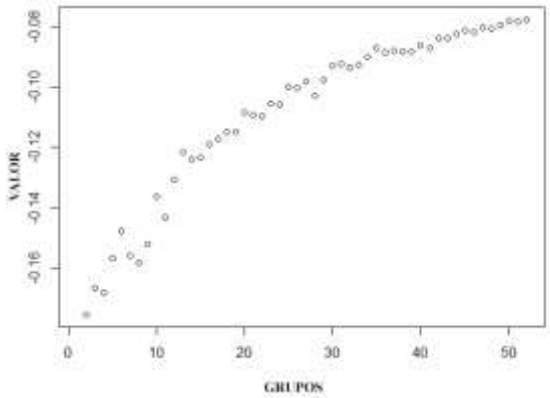

Figura 10 – Índice de validação (POINT BISERIAL)

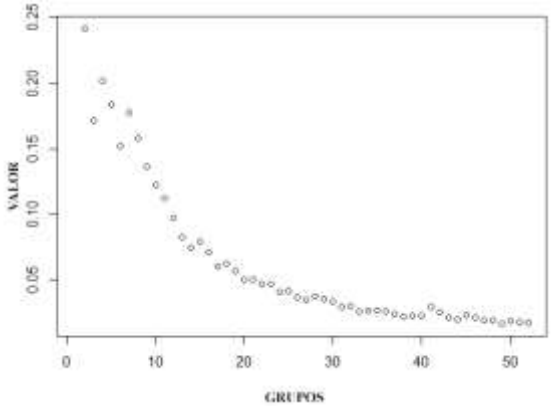

Figura 11 – Índice de validação (PBM)

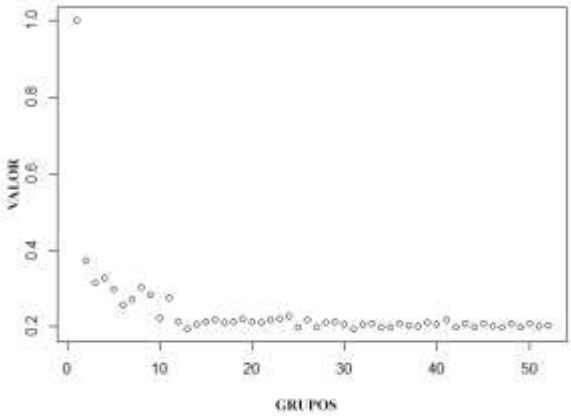

Figura 12 – Índice de validação (Silhueta)

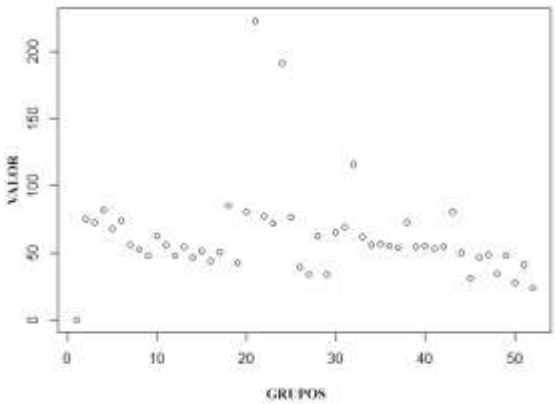

Figura 13 – Índice de validação (XIE BENI)

Pela análise dos gráficos notou-se uma discrepância entre a hipótese de três grupos distintos e os resultados do agrupamento. Pelos índices analisados para o menor valor calculado, o Gamma (Figura 9) estabelece apenas dois clusters pelo menor valor obtido, sendo maior para três grupos ou mais grupos.

O índice Point Biserial (Figura 10) também apresentou um valor de dados crescente, sendo o melhor número de grupos igual a três. O índice Xie Beni (Figura 13) foi o que apresentou o resultado menos conclusivo, pois informou que o número ideal de grupos é de aproximadamente 52.

Para os métodos que indicavam o número de grupos pelo maior valor estatístico (C Index, Calinski Harabasz, PBM e Silhueta) houve, na maioria dos métodos, uma grande preferência por dois grupos de clusters. Novamente, os valores "ótimos" representados por apenas um grupo devem ser desconsiderados.

O índice C Index (Figura 7) indica um comportamento decrescente sendo que o melhor valor válido é de dois grupos. O índice Calinski Harabasz (Figura 9) indica um empate entre dois e três grupos mesmo tendo um comportamento decrescente de valores. O índice Silhueta (Figura 12) apresentou uma tendência para dois grupos apesar de que o valor de referência é extremamente baixo comparado com o valor máximo ideal "1". Apenas o índice PBM (Figura 9) foi o que indicou uma tendência para três grupos.

As análises da Tabela 4 e os resultados obtidos pelos índices de validação demonstram que os grupos 1 e 3 são supostamente o mesmo grupo e que não há diferença significativa.

5 CONCLUSÃO E RECOMENDAÇÕES

Depois de se obter dados em campo, pré-processar, agrupar, validar e analisar os resultados pôde-se concluir que para os dados e os atributos escolhidos, o algoritmo de agrupamento sugere que, para a maioria dos índices avaliados somente dois grupos são necessários para representar os padrões de rugosidade das amostras analisadas, em contraposição aos três grupos propostos pelos autores deste relatório. Tal fenômeno talvez possa ser explicado pelos atributos escolhidos que podem não expressar com exatidão as características presentes nas calçadas.

Outro item que pode ter contribuído para essa conclusão sobre os experimentos de campo foi que a rugosidade do pneu e a pressão nele contida podem interferir na trepidação do veículo e, consequentemente, na captação dos sensores do acelerômetro.

Recomenda-se, para aplicações futuras, um estudo prévio sobre possíveis atributos a serem utilizados para preparação do algoritmo de agrupamento de dados e a utilização de um veículo com menor sensibilidade aos movimentos realizados para se evitar a trepidação desnecessária do equipamento.

Estudos futuros dos participantes desta pesquisa têm como objetivo empregar os resultados deste experimento e utilizá-los em união com variáveis não-quantitativas (nível de segurança, conforto e comodidade) para estipular características que favoreçam a perspectiva do pedestre no quesito de escolha de rotas ou modais para a mobilidade.

6 REFERÊNCIAS

Barbara, D. (2000) An introduction to cluster analysis for data mining. Retrieved November 12, 2003 from <http://www-users.cs.umn.edu/~han/dmclass/clustersurvey 100200.pdf>.

Burden, D. (2001) Building communities with transportation. Transportation research record, Journal of the transportation research board, 1773, 5-20.

Eriksson, J. et al.(2008) The pothole patrol: using a mobile sensor network for road surface monitoring. In: Proceedings of the 6th international conference on mobile systems, applications and services, USA, 17-20 Junho 2008.

Faceli, K., Carvalho, A. C. P. L. F., Souto, M. C. P. (2005) Algoritmos de agrupamento de dados, Relatório técnico de ICMC,249, Universidade de São Carlos, Brasil.

Fiv Asim (2015) Aplicativo de plataforma android androsensor (Versão 1.9.6.3). Disponível em <https://play.google.com/store/apps/details?id=com.fivasim.androsensor &hl=pt_br>.

Fred, A. L. N. (2001) Finding consistent clusters in data partitions. In: J. Kittler & F. Roli (eds.), mcs '01: proceedings of the second international workshop on multiple classifier systems, volume 2096 of lecture notes in computer science, University of Cambridge, UK, 309–318.

González, A., O'Brien, E. J., LI ,Y. Y., and Cashell, K (2008) The use of vehicle acceleration measurements to estimate road roughness. International Journal of Vehicle Mechanics and Mobility, 46 (6), 483–499.

Jain, A. K., Murty, M., Flynn, PJ. (1999) Data clustering: a review. ACM Computing Surveys (csur), 31(3), 264-323.

Mednis, A et al. (2011) Real time pothole detection using android smartphones with accelerometers. In: Distributed computing in sensor systems and workshops (dcoss), International Conference on IEEE, Catalunya, 27-29 Junho 2011,1-6.

Mohan, P., Padmanabhan, V. N., Ramjee, R. (2008) Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In: Proceedings of the 6th acm conference on embedded network sensor systems, USA, 17-20 Junho 2008, 323-336.

The foundation, program and project for statistical computing (2016) (Versão 3.3.2) Disponível em <https://www.r-project.org/>.

Spiric, G. (2014) Algorithm evaluation for road anomaly detection and wear estimation on trucks using an accelerometer, KTH Royal Institute of Technology, Stocolomo, Suécia.