Extracting the relationship between census data and mobile devices

Wandella Maia[1], Mateus P. Silva[1], Fabrício A. Silva[1], Thais R. M. Braga Silva[1]

Instituto de Ciências Exatas e Tecnológicas – Universidade Federal de Viçosa (UFV)
{wandella.maia,mateus.p.silva,fabricio.asilva,thais.braga}@ufv.br

Abstract. In the recent years, we are facing an increasing in the number of mobile users and, consequently, on the amount of data collected from them and available to the industry. Therefore, companies are willing to use such data to discover insights that may help them on providing better and more personalized services. In this work, we explore mobile and census data of thousands of users in Brazil to understand how their local of residence is correlated to their smartphones.

Categories and Subject Descriptors: [**Information systems**]: Data mining

Keywords: census data, correlation, mobile devices

1. INTRODUÇÃO

Nos últimos anos, tem-se percebido um aumento significativo no uso de *smartphones* e, consequentemente, de serviços móveis oferecidos a seus usuários. Os *smartphones*, agindo como sensores, são excelentes fontes de dados sobre os usuários e o ambiente ao qual estão inseridos. Como consequência, as empresas estão cada vez mais interessadas em extrair conhecimento útil desses dados para conhecer melhor os seus usuários e, assim, oferecer serviços melhores e mais personalizados.

Dentre vários aspectos sobre os usuários móveis, podem-se destacar dois: o local de residência e o modelo de aparelho móvel utilizado. Com o local de residência, é possível inferir características socioeconômicas dos usuários, que podem servir como base para direcionamentos de campanhas de marketing e oferta de serviços. Por outro lado, o modelo do *smartphone* e seu respectivo preço pode indicar interesses pessoais e financeiros do usuário.

Com base nesses dois aspectos, surgem duas perguntas:

—Existe alguma correlação entre as características do local de residência do usuário e as características do seu *smartphone*, como por exemplo, se utiliza aparelhos com sistema operacional *Android* ou da *Apple*?
—Existe alguma característica específica do local de residência que pode ajudar a inferir características do *smartphone* do usuário, como por exemplo o preço?

O objetivo deste trabalho é responder a essas perguntas. Para isso, foram utilizados dados reais de mais de 60.000 usuários do Brasil. Foram feitas caracterizações e análises de regressão com os dados. Os resultados revelam que existem algumas características do censo dos locais de residência que podem indicar uma maior probabilidade dos usuários possuírem *smartphones Android* ou da *Apple*. Porém, não foi possível estabelecer com precisão um modelo de regressão para prever o preço médio dos *smartphones* com base em características da região de residência.

Este texto está organizado da seguinte forma: na Seção 2, são discutidos os principais trabalhos que avaliam características demográficas de regiões e de usuários móveis. Em seguida, na Seção 3, os dados utilizados no trabalho são descritos. Os resultados são apresentados e discutidos na Seção 4. Finalmente, as conclusões e trabalhos futuros são listados na Seção 5.

2. TRABALHOS RELACIONADOS

O uso de diversas fontes de dados para a extração de conhecimento sobre o comportamento, características pessoais, mobilidade, interesses e aspectos demográficos tem sido bastante comum nos últimos anos. Com o grande volume de dados publicamente disponíveis, ou coletados por serviços proprietários, as empresas começaram a olhar com atenção para esses dados com o intuito de extrair informações úteis sobre seus clientes, que possam ajudar a melhorar o engajamento e a retenção dos mesmos.

Alguns trabalhos da literatura utilizam as informações de censo coletadas pelos governos para extrair informações e correlações relevantes. [Eagle et al. 2010] avaliaram a correlação entre perfis de uma rede social online e a situação econômica dos usuários com base em regiões censitárias do Reino Unido. [Wesolowski et al. 2013] utilizam dados de migração do censo para aproximar padrões de movimento humano através de escalas temporais. [Remigio et al. 2019] têm como objetivo verificar se a qualidade do bairro pode influenciar na saúde e segurança das pessoas. [De Nadai et al. 2016] têm como objetivo testar a promoção de condição de vida humana em uma cidade da Itália com base no censo.

Outra fonte de informação também muito relevante é referente a dispositivos móveis e sua utilização por meio da telefonia ou acesso à Internet. Esses dados também são explorados por alguns autores em estudos sobre as características dos usuários. [Brdar et al. 2012] tiveram como desafio fazer uma exploração de dados reais de dispositivos móveis para revelar padrões de características demográficas do usuário. O artigo de [Blumenstock et al. 2015] utiliza dados de telefonia móvel de Ruanda, que possui ricas informações, e fornece correlações do uso da telefonia com estatísticas populacionais com agregação regional de censos e pesquisas domiciliares. O artigo de [Smith-Clarke et al. 2014] tem como objetivo analisar registros de assinantes de telefonia móvel e extrair correlações para ajudar o governo a monitorar regiões mais carentes. [Grauwin et al. 2015] tentam detectar padrões de comportamento humano nos principais centros financeiros com base em registros de número de chamadas, *SMS*, transferência de dados e também dados censitários. [Zhong et al. 2013] têm como objetivo criar um *framework* que prevê os dados demográficos dos usuários com base nos dados móveis. O objetivo é utilizar registros de atividades físicas, registros de usuários e contextos ambientais para prever informações como gênero, tipo de emprego, estado civil, dentre outras características.

Além de dados de censos e de dispositivos móveis, outros tipos de dados também têm sido utilizados na análise de perfil de usuários. [Steiger et al. 2015] fazem uma análise espaço-temporal das redes sociais sobre as pessoas e suas atividades sociais em áreas urbanas. Os dados foram obtidos através de serviço de *stream* oferecido pela rede social *Twitter* e para validação, utilizaram dados do censo do Reino Unido. [Lenormand and Ramasco 2016] utilizam aplicações empíricas de *Big Data* para estudar a sistemática de cidades e seus problemas de mobilidade. Com isso, utilizou-se 20 milhões de *tweets* de vários lugares do mundo para quantificar e comparar as 58 áreas mais influentes no mundo, levando em conta a área coberta pelos usuários. [Bogomolov et al. 2014] utilizam dados sobre o comportamento humano agregado e anônimo, derivados da atividade da rede móvel, para lidar com o problema da previsão de crime. Sendo assim, foram utilizados para análise três conjuntos de dados: um com a localização de 124.119 pessoas de Londres, um com todos os crimes reportados do Reino Unido e o último com as métricas da população de Londres.

Diferente dos trabalhos existentes, o presente estudo analisa dados de residência dos usuários, com suas respectivas características demográficas, e o aparelho móvel que o usuário possui. O objetivo é investigar se existe alguma correlação entre esses dois aspectos, e se essa correlação pode ser usada para se conhecer melhor o perfil do usuário. Esta é a primeira vez que esse tipo de correlação é feita com uma amostra real com dezenas de milhares de usuários, e com informações detalhadas do setor de residência.

3. OS DADOS

3.1 Usuários

Para a realização deste trabalho, foi disponibilizada por uma empresa parceira que atua na área de serviços móveis uma amostra de dados com 64.108 usuários, contendo a localização aproximada da residência do usuário e o modelo do *smartphone* utilizado. Os dados foram coletados com o consentimento dos usuários, e disponibilizados para a pesquisa sob as penas de um termo de confidencialidade assinado entre as partes. Esses usuários estão presentes em 3.665 municípios do Brasil, e distribuídos no mapa conforme mostra a Figura 1.

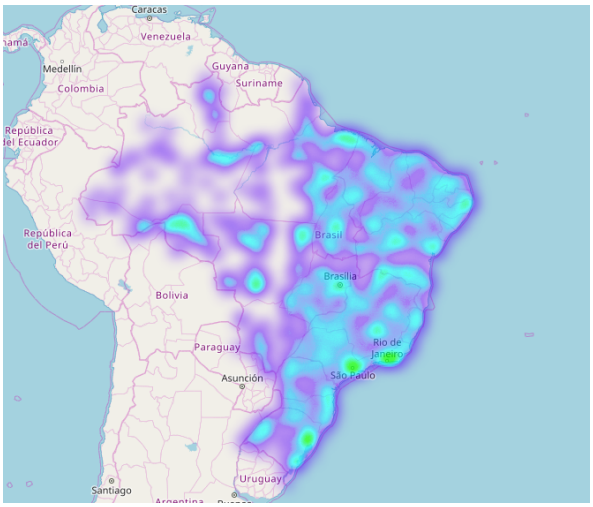


Fig. 1. Mapa de distribuição dos 64.108 usuários ao longo do Brasil.

3.2 Características dos *Smartphones*

A base de usuários possui apenas um nome técnico indicando o modelo do dispositivo móvel e coletar dados sobre as características de cada um manualmente tornaria o trabalho inviável. Portanto, foi utilizado uma ferramenta chamada CrawMobi para complementar a base do trabalho.

3.2.1 *CrawMobi.* O CrawMobi é um web service que permite a comunicação entre aplicações na internet e o objetivo principal desta ferramenta é auxiliar o usuário a pesquisar informações técnicas sobre smartphones. Sendo assim, existem conceitos e técnicas aplicados nele como recuperação da informação, mecanismos de busca e web service que fazem dele um sistema bom e robusto.

A sua recuperação da informação é realizada através de documentos web, trabalhando no nível de recuperação sobre o tipo de dados. Logo, sua busca está voltada para o local do conteúdo e que é baseada em dados específicos, o que o torna mais eficiente.

Para fazer a recuperação da informação é necessário um mecanismo de busca eficiente e para isso foi utilizado um crawler. Ele é responsável pela descoberta, análise e indexação do conteúdo web. Sendo assim, ele vasculha a rede em busca de diversos documentos com base na entrada de parâmetros e após descobrir um potencial documento, ele captura seu conteúdo e insere o seu endereço eletrônico em uma lista de links úteis encontrados. O processo de análise se dá quando a página é capturada e a ferramenta recupera as informações da tag HTML. Por fim, é realizado a indexação do conteúdo com os links salvos.

O CrawMobi é composto por vários métodos que podem ser usadas por outras aplicações. Tais métodos visam permitir a transmissão de dados utilizando protocolos de comunicação de rede de diferentes aplicações independentemente de tecnologias e ferramentas utilizadas para sua criação. Ele também usa a arquitetura REST(*Representational State Transfer*) em sua implementação, onde REST, em português, Transferência de Estado Representacional, como os outros *Web Services*, fornece interoperabilidade entre sistemas na Internet. Esta arquitetura possui um conjunto de restrições e propriedades definidos para comunicação via HTTP. Os sistemas que solicitam o serviço podem acessar e manipular recursos da Web de padronização e encapsulamento de dados (representados por textos, JSON, XML e outros), utilizando um conjunto de operações sem estado. Além disso, tal arquitetura foi escolhida porque a mesma é muito utilizada no mercado e por apresentar facilidade em recuperar a informação desejada.Isso faz com que o cliente consuma os dados dos servidores de maneira mais rápida, baixando somente o pacote requisitado.

Na figura 2 apresenta a arquitetura do Crawmobi,contendo todos os seus componentes e o fluxo da informação desde a requisição do cliente até o retorno com a caracterização dos dados. A interação com a aplicação é através de uma interface. Primeiramente, é necessário criar um arquivo excel com os nomes comerciais dos aparelhos desejados para a coleta, descritos linha após linha na primeira coluna do arquivo. Os nomes comerciais devem ser compostos pela marca do aparelho seguido de seu modelo. Após isso, deve-se fazer o *upload* do arquivo na página do Web Service. Logo, o serviço retorna um arquivo contendo 22 características dos aparelhos listados.

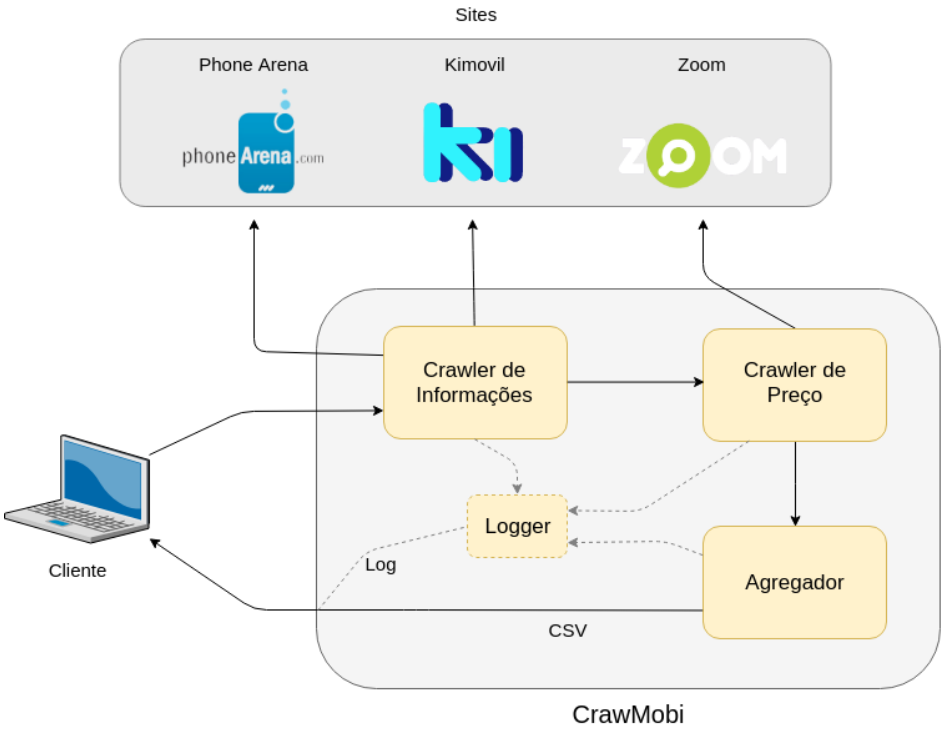
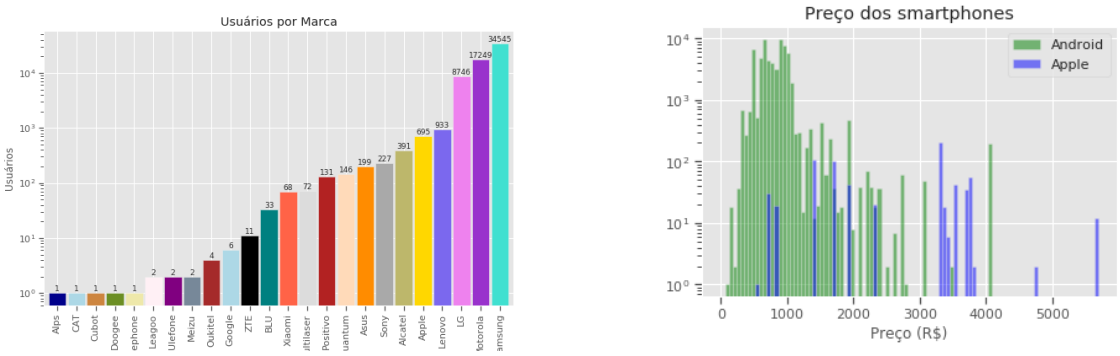


Fig. 2. Arquitetura do CrawMobi

3.3 Setor Censitário

Os dados com os setores censitários foram obtidos através do *site* governamental do IBGE (Instituto Brasileiro de Geografia e Estatística) com o censo do ano de 2010. O IBGE é um instituto com o intuito de realizar pesquisas sobre a população brasileira e esses dados buscam detalhar as condições



(a) Marcas utilizadas por usuários

(b) Histograma de preços

Fig. 3. Características dos *smartphones* dos usuários, sendo o histograma dos preços e a quantidade de usuários por marca.

em que a população vive. O setor censitário é uma área menor (existem vários setores por cidade) de controle onde o objetivo é saber informações cadastrais daquela área. O arquivo no formato *shapefile* espacial foi obtido e contém polígonos com toda a divisão dos 3.665 municípios cobertos pelos dados, totalizando 310.120 setores. O uso de setores censitários e não de cidades é um diferencial deste trabalho, pois um setor representa mais precisamente as características demográficas, com poucas variações entre seus residentes.

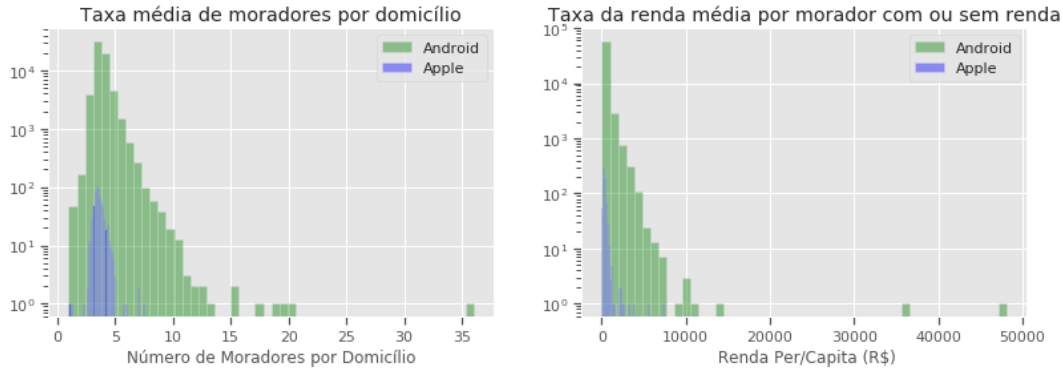
4. ANÁLISE E RESULTADOS

Inicialmente, para responder à primeira pergunta, foi feita uma caracterização dos dados para que seja possível extrair conhecimentos sobre a relação entre as características do local de residência com as características do *smartphone*, em especial se são aparelhos *Android* (de várias marcas) ou *iOS* (*Apple*). Em seguida, foi feita uma tentativa de estimar um regressor linear para prever o preço médio dos *smartphones* de um setor censitário com base em suas características, com o objetivo de responder à segunda pergunta levantada.

4.1 Caracterização

A Figura 4(a) apresenta como estão distribuídos os *smartphones* pelo número de moradores por domicílio, diferenciando-os pelo sistema operacional. Como o número de aparelhos com iOS é inferior ao de *Android* (ver Seção 3.2), é esperado que a curva descrita no gráfico para o sistema operacional da *Apple* esteja abaixo da do concorrente, o que de fato ocorre, e que o comportamento das duas sejam próximos, o que não acontece. De fato, setores censitários com média de cinco ou mais moradores por domicílio não apresentam o número de *smartphones* com *iOS* que deveriam, o que mostra uma correlação negativa entre essas variáveis. Ou seja, é possível supor que quanto mais moradores há em uma residência, menores são as chances de que algum deles tenha um celular com *iOS*, e que usuários da *Apple* normalmente dividem sua residência com outras poucas pessoas.

A Figura 4(b) apresenta as distribuições de renda per capita dos setores em termos do sistema operacional dos usuários. Apesar de os aparelhos da *Apple* serem em geral mais caros (como mostra a Figura 3), não foi possível observar que os usuários *Apple* estejam em regiões com renda per capita maior. Apesar de termos uma amostra menor desses usuários, ainda assim era esperado um comportamento diferente. Entretanto nota-se claramente que os usuários de *Apple* estão em uma faixa de renda per capita menor. Esse comportamento pode ser justificado pela existência de pessoas com renda superior à média do setor censitário em que reside, pela possibilidade de parcelamento na compra de *smartphones*, ou pelo mercado de aparelhos usados a preços mais acessíveis. No entanto,

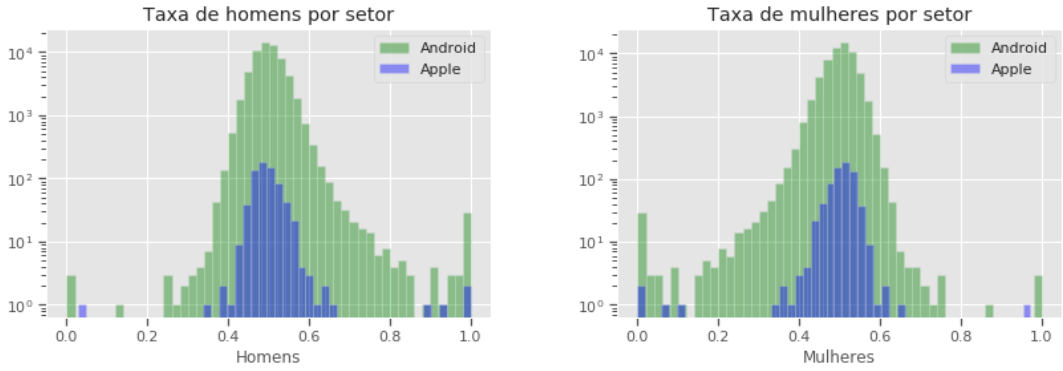


(a) Moradores por domicílio. (b) Renda per/capita.

Fig. 4. Histogramas do número de moradores por domicílio e da renda per/capita.

não é possível comprovar nenhuma dessas hipóteses com os dados utilizados neste trabalho.

Em relação ao gênero, a Figura 5 mostra que em setores censitários em que a taxa de pessoas do gênero masculino é maior, a proporção de usuários *Apple* é menor. Ou seja, em locais com uma predominância um pouco maior de mulheres, a proporção de usuários *Apple* é maior.



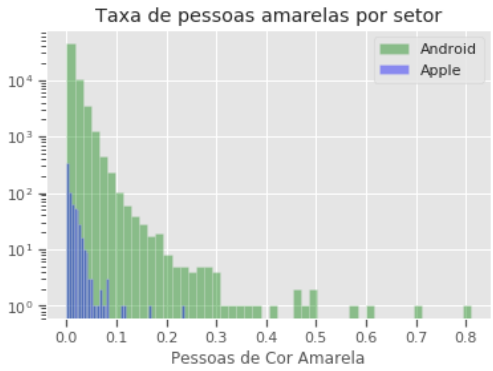
(a) Histograma do gênero masculino (b) Histograma do gênero feminino

Fig. 5. Histograma da taxa média de (a) homens e (b) mulheres

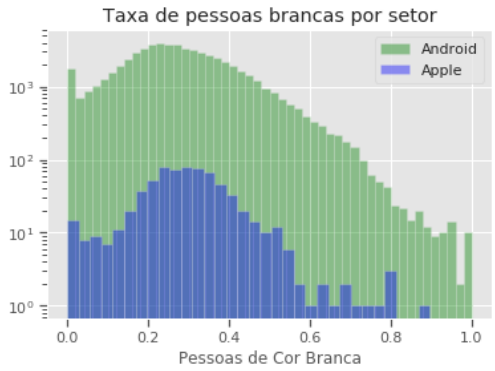
Em relação às características de cor e raça, pode-se perceber na Figura 6 que há muita diversidade nas curvas apresentadas. Dentre esses resultados, pode-se destacar que em setores em que a taxa de negros é menor, maior é a proporção de usuários *Apple*. Por exemplo, não foi registrado nenhum usuário *Apple* em setores com mais de 50% de negros. O mesmo é válido para amarelos, sendo que a maior proporção de usuários *Apple* está nos setores com predominância parda ou branca.

Por fim, em termos dos atributos de idade e alfabetização pode-se observar outras características interessantes. A Figura 7(a) mostra que quanto menos alfabetizado for o setor, maior é a proporção dos usuários *Android*. Além disso, na Figura 7(b) percebe-se que os usuários *Android* ocupam uma população de faixas etárias mais abrangentes, enquanto os usuários *Apple* ocupam uma faixa de população mais jovial entre 20 e 40 anos.

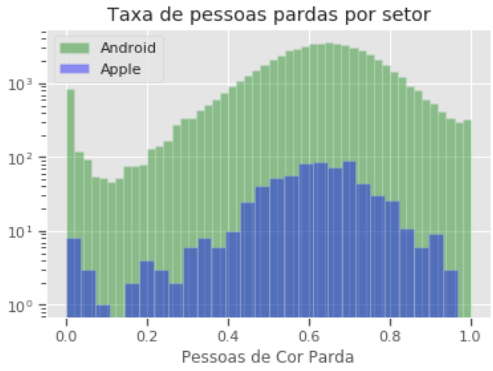
Com base nos resultados obtidos, é possível perceber que existe alguma correlação entre as características de uma região e a proporção de usuários *Apple* e *Android* dessa região. Em resumo, foram observadas diferenças nas proporções para características de idade, alfabetização, gênero, número de



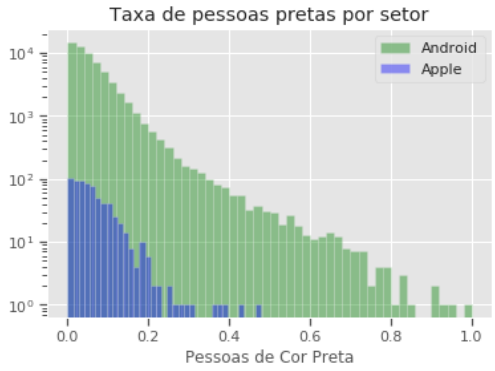
(a) Histograma de pessoas de cor amarela



(b) Histograma de pessoas de cor branca



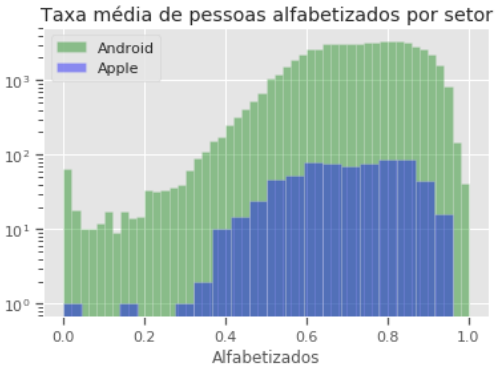
(c) Histograma de pessoas de cor parda



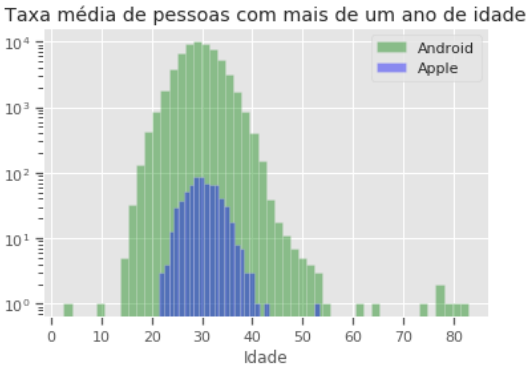
(d) Histograma de pessoas de cor preta

Fig. 6. Histograma em termos de cor/raça

moradores e cor. Diferentemente do esperado, a renda per capita do setor não foi um fator relevante, mesmo considerando que aparelhos da marca *Apple* são mais caros que marcas que adotam *Android* em geral.

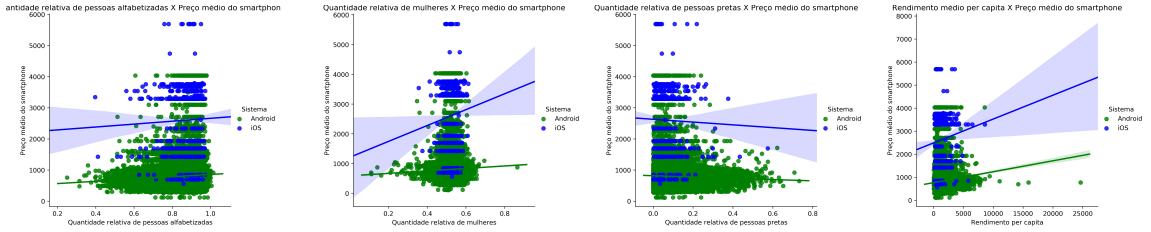


(a) Histograma de alfabetizados



(b) Histograma média de idades

Fig. 7. Histograma da taxa média de (a) alfabetizados e (b) idades



(a) Alfabetizados. Androids: 0.1. iOS: 0.033

(b) Mulheres. Androids: 0.039. iOS: 0.073

(c) Pretos. Androids : −0.048. iOS: −0.024

(d) Renda per capita. Androids: 0.01. iOS: 0.088

Fig. 8. Correlações analisadas em relação ao preço do smartphone com coeficiente de Pearson para smartphones com Android e iOS explicitados nas legendas

Atributos usados	R^2	**Validação cruzada**	**Desvio médio quadrático**
idade/mulheres/pretos/amarelos	0.02	-0.04	381.10
moradores domicílio/mulheres/pretos/amarelos	0.02	-0.04	381.13
mulheres/pretos/amarelos	0.02	-0.03	381.13
idade/moradores domicílio/mulheres/amarelos	0.02	-0.05	381.14
idade/mulheres/amarelos	0.02	-0.04	381.14
moradores domicílio/mulheres/amarelos	0.02	-0.04	381.16
mulheres/amarelos	0.02	-0.03	381.16
idade/moradores domicílio/pretos/amarelos	0.02	-0.04	381.21
idade/pretos/amarelos	0.02	-0.03	381.22

Table I. Resultados para diferentes combinações de atributos do setor censitário como entrada.

4.2 Análise de Regressão

Para responder à segunda pergunta, foram utilizados modelos de regressão linear com o objetivo de estimar o preço médio dos *smartphones* de um setor censitário com base em suas características. Neste trabalho, aplicou-se diferentes combinações de atributos da base de dados de setores censitários como variáveis explicativas. Essas combinações foram realizadas inicialmente em pares, e em seguida foram feitas várias alternativas com potencial de bons resultados.

Na Figura 8, são apresentados os principais resultados de correlação par-a-par. Os atributos do setor censitário referentes à taxa de alfabetização, taxa de mulheres, taxa de pessoas negras e rendimento per capita foram os que tiveram melhores correlações com o preço médio dos *smartphones* do setor. Porém, é possível perceber que, pelo menos visualmente, a correlação não é clara. Essa falta de correlação é válida tanto para aparelhos *Android* quanto *Apple*.

Para verificar se uma combinação com mais variáveis explicativas pudesse estar correlacionada ao preço do *smartphone*, foram criados diferentes modelos de regressão linear multi-variada, alternando os atributos dos setores utilizados como entrada. A tabela I mostra as nove combinações de variáveis explicativas que tiveram os menores erros. Pode-se perceber que mesmo essas combinações tiveram um valor de R^2 muito baixo, o que mostra que o preço não é explicado pelas variáveis na grande maioria das vezes. Além disso, os erros em geral foram significativamente altos, como é possível perceber na coluna de desvio médio quadrático, e o algoritmo de validação cruzada também apresentou resultados ruins. Tudo isso aconteceu independente da combinação de atributos usada, o que reforça a tese de que essas variáveis, em geral, não apresentam um grau de significância suficiente com esse conjunto de dados para a previsão do preço do *smartphone*. Em resumo, não foi possível encontrar um modelo para estimar o preço médio dos *smartphones* dos usuários com base nas características do local de residência dos mesmos.

5. CONCLUSÃO

Este trabalho teve como objetivo avaliar possíveis correlações entre características demográficas do setor censitário de residência e as características do *smartphone* de um grande volume de usuários móveis. Para isso, foram feitas caracterizações em termos de renda, cor/raça, gênero, alfabetização, idade e número de moradores. Como resposta à primeira pergunta levantada, foi possível verificar que algumas características do setor fazem com que a proporção de usuários *Apple* seja diferente dos usuários *Android*.

Após a caracterização, foram feitas análises com modelos de regressão com o objetivo de responder à segunda pergunta levantada e prever o preço médio de *smartphones* de um setor censitário com base em suas características. Ao se tentar várias combinações de variáveis explicativas, verificou-se que não existe uma correlação clara e, consequentemente, não foi possível criar um modelo de regressão preciso para prever o preço dos *smartphones*.

Este trabalho é importante porque atualmente as empresas têm buscado oferecer melhores serviços a seus usuários. Portanto, esta análise pode auxiliá-las a direcionar as ofertas e promoções de serviços e produtos ao público-alvo que potencialmente terá interesse. Como trabalhos futuros, pretende-se avaliar outros modelos de aprendizagem para estimar o preço e o sistema operacional dos usuários. Além disso, pretende-se também fazer análises em relação às características das cidades e estados.

REFERENCES

Blumenstock, J., Cadamuro, G., and On, R. Predicting poverty and wealth from mobile phone metadata. *Science* 350 (6264): 1073–1076, 2015.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI '14. ACM, New York, NY, USA, pp. 427–434, 2014.

Brdar, S., Culibrk, D., and Crnojevic, V. Demographic attributes prediction on the real-world mobile data. In *Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK*, 2012.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., and Lepri, B. The death and life of great italian cities: a mobile phone data perspective. In *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, pp. 413–423, 2016.

Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science* 328 (5981): 1029–1031, 2010.

Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., and Ratti, C. Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational approaches for urban environments*. Springer, pp. 363–387, 2015.

Lenormand, M. and Ramasco, J. J. Towards a better understanding of cities using mobility data. *Built Environment* 42 (3): 356–364, 2016.

Remigio, R. V., Zulaika, G., Rabello, R. S., Bryan, J., Sheehan, D. M., Galea, S., Carvalho, M. S., Rundle, A., and Lovasi, G. S. A local view of informal urban environments: a mobile phone-based neighborhood audit of street-level factors in a brazilian informal community. *Journal of Urban Health*, 2019.

Smith-Clarke, C., Mashhadi, A., and Capra, L. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp. 511–520, 2014.

Steiger, E., Westerholt, R., Resch, B., and Zipf, A. Twitter as an indicator for whereabouts of people? correlating twitter with uk census data. *Computers, Environment and Urban Systems* vol. 54, pp. 255–265, 2015.

Wesolowski, A., Buckee, C. O., Pindolia, D. K., Eagle, N., Smith, D. L., Garcia, A. J., and Tatem, A. J. The use of census migration data to approximate human movement patterns across temporal scales. *PLOS ONE* 8 (1): 1–8, 01, 2013.

Zhong, E., Tan, B., Mo, K., and Yang, Q. User demographics prediction based on mobile data. *Pervasive and Mobile Computing* 9 (6): 823 – 837, 2013. Mobile Data Challenge.