Um Padrão de Agrupamento de Cidades a Partir da Dinâmica Social Urbana Extraída de Dados Provenientes de Redes de Sensores Participativos

Vilker T. C. Lobo[1], **Thiago H. Silva**[2], **Heitor S. Ramos**[1]

[1]Instituto de Informática – Universidade Federal do Alagoas – UFAL, Brasil

[2]Universidade Tecnológica Federal do Paraná - UTFPR, Brasil

vilker.tenorio@gmail.com, heitor@ic.ufal.br, thiagoh@utfpr.edu.br

Abstract. *The social dynamics of cities have been studied over the years, however, with the popularization of the use of smartphones, massification of the Internet and social networks, a new form of such study has emerged. In this context, this paper presents a new way of comparing cities, using the mobility pattern of its inhabitants as a measure of the city's similarity. To validate the study,* 1.805.029 *Foursquare check-ins were distributed across* 8 *cities, with the purpose of - using the proposed metrics - presenting different patterns of city grouping.*

Resumo. *A dinâmica social das cidades vem sendo estudada ao longo dos anos, no entanto, com a popularização do uso dos smartphones, massificação da internet e das redes sociais, tem surgido uma nova forma de se realizar tal tipo de estudo. Nesse contexto, este trabalho apresenta uma nova maneira de se comparar cidades, utilizando como medida de similaridade o padrão de mobilidade dos seus habitantes. Como modo de validar o estudo, foram utilizados* 1.805.029 check-ins *do* Foursquare *distribuídos por* 8 *cidades, a fim de - utilizando as métricas propostas - apresentar diferentes padrões de agrupamento das cidades.*

1. Introdução

O estudo do comportamento populacional é uma área de vital importância para a sociedade e vem sendo realizado desde os meados do século XIII, quando *D. Afonso III*, rei de *Portugal*, procedeu uma contagem populacional para saber quantos homens estavam aptos à guerra. Desde então, houve uma enorme evolução tecnológica e as estatísticas populacionais passaram a ser realizadas de modo cada vez mais eficiente. Contudo, na grande maioria dos casos, tais estudos são feitos por meio de entrevistas e formulários respondidos pela população, o que, conforme [Kostakos et al. 2009], implica grande consumo de recursos e tempo, dificulta a coleta longitudinal de dados e possibilita que os dados sejam influenciados por erro de autorelato.

Como alternativa para os problemas apresentados, o uso de dados de redes sociais tem se popularizado nos últimos anos e possibilitado uma nova forma de realizar estudos populacionais, já que essas redes passam a armazenar um grande volume de dados de seus usuários à medida que são utilizadas.

Segundo matéria da [Stewart 2016], os usuários passam, em média, 50 minutos por dia conectados as redes sociais, isso considerando apenas as plataformas *Facebook*, *Messenger* e *Instagram*, sem falar nas ferramentas, como *Youtube*, *Qzone*, *WhatsApp*, *FourSquare* e *Google+*. Esse valor pode até não parecer tão elevado, mas basta considerar que ele corresponde a 6% do tempo útil diário e que supera quase todas as outras atividades de lazer, se aproximando, até mesmo, do tempo gasto em necessidades básicas como comer e beber, 1,07 horas diárias.

Usualmente agrupa-se a população por parâmetros como o comportamento socioeconômico [Tammaru et al. 2016], religioso [Kosmin and Lachman 2011] e étnico [Waters 1994]. A utilização de dados de redes sociais permite não somente a análise comportamental que tradicionalmente se vê nas pesquisas, mas também o estudo da dinâmica social da população por meio de suas atividades diárias. Verificando as semelhanças e diferenças dessas atividades, pode-se agrupar regiões pelo seu padrão de mobilidade urbana, no entanto, realizar tal estudo é uma tarefa árdua, dado que é necessário acompanhar a dinâmica diária da amostra estudada.

Alguns trabalhos têm abordado temas que usam as Redes Sociais como fonte de dados. Por exemplo, [Frias-Martinez et al. 2012] e [Phithakkitnukoon and Olivier 2011] utilizando tais dados para inferir as atividades existentes em determinadas regiões das cidades, em especial, [Cranshaw et al. 2012] estuda as características das cidades dividindo-as em *Livehoods* que é uma nova forma de dividir a cidade pelas atividades predominantes em cada região, para isso o autor analisa 18 milhões de postagens feitas no *Foursquare* e valida seus resultados comparando os grupos formados na cidade de Pittsburgh com a opinião de 27 moradores da cidade.

Outros autores utilizam dados de Redes Sociais juntamente com dados de sensores físicos para obter uma maior precisão nos resultados, como o trabalho de [Xie and Wang] que associa 125 milhões de dados veiculares capturados por sensores de tráfego com 3,4 milhões de postagens - provenientes do *Twitter*, *Foursquare*, *Flickr*, *Picasa*, e *Panoramio* - com intuito de estudar possíveis soluções para a melhoria da mobilidade urbana.

Estudos recentes visam entender o padrão de mobilidade urbana, como [Noulas et al. 2011], [Silva et al. 2012a] e [Silva et al. 2012b]. Neste grupo podemos destacar [Silva et al. 2014] que usa observações capturadas do *Twitter* com a tag do *Foursquare* contendo as 3 seguintes categorias: *Drink*, *Fast-Food* e *Slow-Food* para estudar a preferência dos usuários em relação a estes 3 aspectos em diferentes locais do mundo. Para validar os resultados obtidos, o autor compara seus resultados com um *survey* [Inglehart and Welzel 2010] que estudou aspectos religiosos, políticos, econômicos e de estilo de vida da população de alguns países no período período de 2005 à 2008.

O presente trabalho também objetiva estudar a dinâmica social urbana das cidades, no entanto, difere de outros, como [Silva et al. 2012b], [Silva et al. 2012a] e [Silva et al. 2014], pois oferece uma investigação mais aprofundada das particularidades dos dados. Especificamente, mostra-se que o comportamento social de um dia da semana em particular difere dos demais dias, enquanto os trabalhos citados não fazem esta distinção. Além disso, realiza-se um estudo mais abrangente, investigando mais vertentes da dinâmica urbana (estudando cada característica do comportamento social da população separadamente), e propõe-se um modelo matemático mais consistente representando os

dados por Cadeias de Markov e utilizando robustos testes estatísticos para comparar diferentes cidades.

2. Sensoriamento Participativo

Sensoriamento Participativo é um tipo de sensoriamento remoto onde pessoas agem como sensores, transmitindo, de forma voluntária, suas sensações sobre o ambiente em que se encontram. Um grupo de indivíduos, atuando em conjunto, formam uma *Redes de Sensores Participativos (RSP)*, onde cada um passa a atuar como um nó móvel da rede, gerando informações que, uma vez reunidas, formam uma base de conhecimento sobre o fenômeno que está sendo observado.

Sistemas que utilizam dados de *SP* para seu funcionamento são chamados de *Sistemas de Sensoriamento Participativo (SSPs)*. Estes, alavancados pela massificação do uso dos smartphones e outros dispositivos conectados em rede, se tornaram muito populares nesta década, notadamente por mesclarem aspectos de redes sociais e mecânicas de jogo em seu funcionamento.

Os *SSPs*, comumente conhecidos como redes sociais, geralmente se baseiam e refletem as relações sociais da vida real por meio de plataformas online, como um site, nas quais os usuários podem compartilhar ideias, atividades, eventos e interesses por meio da internet [Zheng and Zhou 2011].

Com a evolução da internet móvel - que, segundo [ICT Data and Statistics Division 2016], oferece cobertura banda larga à 84% da população mundial, com 67% de cobertura em áreas rurais - as pessoas passaram a estar conectadas por meio de seus dispositivos móveis em todos os momentos do dia. Essa interação constante de cada usuário com a rede deixa um rico rastro de dados nos *SSP's* que a medida que os acumula passam a atuar como grandes *RSP's*.

É importante levar em consideração, no entanto, que atividades de sensoriamento descritas pelo usuário são passíveis ao erro, seja por conta de má utilização da ferramenta – como quando um usuário partilha uma foto qualquer afirmando ser do local em que se encontra - ou por conta do estado momentâneo do ambiente – quando, por exemplo, devido a um feriado uma determinada loja está fechada e o usuário indica erroneamente que o estabelecimento sempre está fechado nesse dia da semana. Visando evitar tais problemas, os *SSPs* implementam várias medidas de proteção, como algoritmos de análise de imagens, análise textual, além de rankeamento de usuários e políticas de uso restritivas, que provocam em alguns casos exclusão de conteúdo e até mesmo, em casos extremos, banimento de contas que não se adequem à política de uso da ferramenta.

Além disso, por não haver, por parte do usuário, a obrigação de fornecimento a informação, nem mesmo de manter o dispositivo operante, o trabalho com dados de *RSP* pode se tornar uma tarefa desafiadora, pois, ao contrário de redes de sensores convencionais, não há garantias da periodicidade da informação disponibilizada. Tal característica deve sempre ser levada em consideração na construção dos experimentos que utilizem esse tipo de fonte de dados.

Contudo, apesar dos problemas apresentados, o uso de dados de *RSP's* ainda apresenta grandes vantagens como o custo e a abrangência da informação, principalmente se comparadas com redes tradicionais uma vez que existe um alto valor envolvido na

implementação redes de sensores tradicionais para sensoriamento em larga escala. Ao se utilizar dados de *RSP's* é possível o acesso a um grande volume de dados tendo como único custo a coleta dos mesmos, já que os dispositivos da rede funcionam de maneira autônoma e não apresentam problemas comuns em redes sem fio tradicionais, como vida útil do dispositivo e gasto de energia.

3. Caracterização dos Dados

Para este trabalho, foram utilizados dados de um *PSS* que oferece serviços de compartilhamento de localização, o *Foursquare*. Segundo [Noulas et al. 2011], este tipo de serviço é construído baseando-se na noção de aproximar locais que visitamos com os amigos aos quais estamos conectados. Devido a sua popularidade crescente, oferecem uma base de dados de atividade humana promissora.

O *Foursquare* é um dos mais populares *SSP's* que oferecem serviços de localização, e sua plataforma é dividida em duas aplicações, o *Foursquare* e o *Swarm*, sendo acessível via navegador e possuindo versões de aplicativo para as principais plataformas móveis (*iOS*, *Android*, *Windows Phone*, *Blackberry* e *Symbian*). A divisão da plataforma se dá pela foco de cada aplicação, enquanto o *Swarm* é voltado para *check-ins* e interação entre usuários, o *Foursquare* se dedicada a descobertas e recomendações de locais. Segundo [Foursquare 2016] possui:

- Comunidade: mais de 50 milhões de pessoas usam o *Foursquare* e o *Swarm* a cada mês, em desktops, internet móvel e aplicativos móveis. Pessoas fizeram check-in mais de 8 bilhões de vezes mundialmente.
- Plataforma: mais de 65 milhões de locais formam o mapa de negócios no mundo.

3.1. Extração dos Dados

Existem diversas formas de coleta de dados de *PSS's*, as mais comuns são por meio de *API's*, *Web Crawler* e aplicações. Tais processos nem sempre são simples, pois podem demandar bastante tempo, processamento e em alguns casos até um custo financeiro.

Dentre as *API's*, existem dois tipos mais usados para coleta de dados, são elas *REST* e *STREAMING*. Em ambos os casos existem limites severos de quantidade de dados que podem ser extraídos. Tais limites foram criados pois uma das fontes de renda destas ferramentas é a venda de dados para terceiros que desejem minerá-las.

No caso específico do *Foursquare*, sua *API* fornece uma quantidade muito limitada de dados de forma gratuita, como podemos ver em [Foursquare 2017]. Com tudo é possível utilizar outro canal para adquirir tais dados, que é por intermédio de *check-ins* do *Foursquare* que são também compartilhadas no *Twitter*. Baseado no trabalho de [Noulas et al. 2011], é esperado que amostras coletadas utilizando tal mecanismo correspondam de 20% à 30% da quantidade total de dados compartilhados pela plataforma.

Neste trabalho em especial, em virtude da alta demanda computacional e de rede necessária, foram utilizados dados coletados e disponibilizados, por [Silva 2014] que também foram utilizados em [Silva et al. 2012a] e [Silva et al. 2014]. Nesse conjunto de dados estão armazenados 33 dias do fluxo de *check-ins* do *Foursquare* compartilhadas no *Twitter*, no periodo entre os dias 25-04-2014 à 18-06-2014. Dentre esses dias ocorreram interrupções na captura de dados - dos dias 04-05-2014 à 06-05-2014, 20-05-2014 à

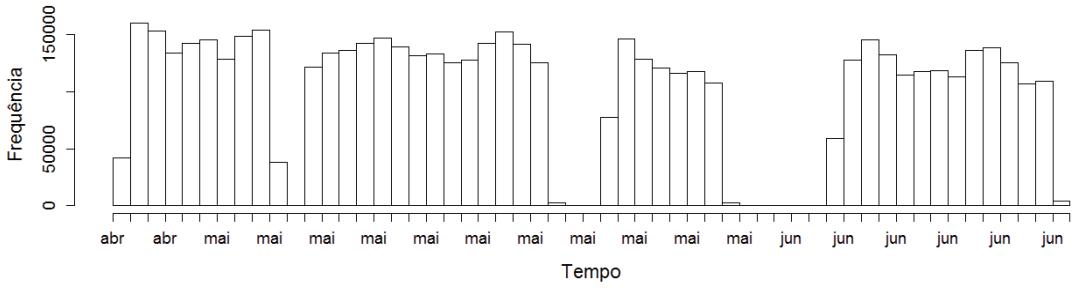

Figura 1. Histograma da frequência de *check-ins*, de 25-04-2014 à 18-06-2014

22-05-2014 e 30-05-2014 à 04-06-2014 – sendo possível vê-las, bem como a frequência diária de *check-ins* em todo o periodo de extração de dados, na figura 1.

Um *Check-In* é muito rico em informações, contendo dentre outros parâmetros, geolocalização, tempo exato de sua postagem, além de um link para uma página do *Foursquare* que possui informações sobre o local onde foi realizada. Utilizando tal link é possível, recorrendo a técnicas de *Web Crawler*, extrair desta página informações sobre a classificação do local da postagem. Esta técnica foi também fui utilizado por [Noulas et al. 2011], [Phithakkitnukoon and Olivier 2011] e [Silva et al. 2014] em seus trabalhos.

3.2. Tratamento dos Dados

O *Foursquare*, conforme visto anteriormente, tem um vasto banco de dados de locais cadastrados, tais locais são organizados de maneira hierárquica por meio de categorias. Neste trabalho iremos utilizar apenas com 9 das 10 categorias primárias do *Foursquare*, são elas: *Arts & Entertainment (A&E), College & University (C&U), Food (FD), Nightlife Spot (NL), Outdoors & Recreation (O&R), Professional & Other Places (P&O), Residence (RD), Shop & Service(S&S)* e *Travel & Transport (T&T)*. Todos os locais classificados por categorias que são subcategorias destas foram reclassificadas como a categoria principal. A exclusão da décima categoria, *Event*, se deu por conta da não periodicidade da ocorrência de *check-ins* na mesma, já que esta caracteriza atividades esporádicas e que em geral não se repetem, ou se repetem com um período maior do que a amostra que iremos utilizar nesse trabalho, como é o caso de festivais de música, conferências e convenções.

Ao término da etapa de coleta de dados no *Twitter* e a posterior busca das categorias no *Foursquare*, foram escolhidas 10 cidades (New York, Bangkok, Los Angeles, Manila, Mexico City, Moscow, Recife, Rio de Janeiro, São Paulo, St. Petesburg) para, com o auxilio dos dados de geolocalização contidos em cada *Check-In*, criar um subconjunto de dados com postagens realizadas apenas nas cidades selecionadas. Para isso foram cruzadas as localizações de cada mensagem compartilhada com os *Bounding Box* das cidades – caixas imaginárias que definem os limites geográficos das cidades em termos de latitude e longitude – resultando em um novo conjunto de dados apenas com dados das cidades selecionadas.

Após os dados terem sido separados por cidade, foram então separados por dia, no entanto não foi escolhido o formato tradicional de representação do dia, que começa à meia noite e vai até á meia noite do outro dia, foi escolhido trabalhar com o formato

de dia social que vai das 4 às 4 da manhã. Essa abordagem também é utilizada em outros trabalhos, como [Silva et al. 2012b], e é feita principalmente por conta da categoria *Nightlife Spot*, que representa locais de vida noturna, comumente frequentados até muito após à meia noite. Vale salientar que como muito destes locais também funcionam pelo dia, como alguns bares e casas de shows por exemplo, então todas as publicações feitas nestes locais fora do horário noturno, 18 às 4, foram desconsideradas.

Como o foco desse trabalho é em estudar a dinâmica das cidades, foram considerados apenas *check-ins* de usuários com uma certa periodicidade. Mais especificamente, foram apenas consideradas postagens feitas por um mesmo usuário em um mesmo dia social e com até no máximo 6 horas de diferença e com diferença mínima de 10 minutos se as categorias forem diferentes e 30 se forem iguais. Tal escolha foi feita por assim existir uma maior probabilidade destas postagens representarem o real padrão de mobilidade do usuário.

Já que o conjunto de dados conta apenas com *check-ins* feitos por usuários com uma frequência diária de postagens e os locais onde onde tais *check-ins* foram realizados estão todos classificados por categorias do *Foursquare*. Foi possível então, agrupando as amostras por usuário e data, construir uma sequência de transições de tipos de locais por onde os usuários esteveram ao longo de um determinado dia. É possível ver um exemplo dessa sequência na tabela 1.

Tabela 1. Exemplo da mobilidade diária de um usuário na cidade de Los Angeles.

IDUSER	TIMESTAMP-BEGIN	TIMESTAMP-END	CATEGORY-BEGIN	CATEGORY-END
27377355	2014-05-18 09:28:33	2014-05-18 10:36:23	Food	Residence
	2014-05-18 10:36:23	2014-05-18 12:34:25	Residence	Arts and Entertainment
	2014-05-18 12:34:25	2014-05-18 13:20:23	Arts and Entertainment	Residence
	2014-05-18 13:20:23	2014-05-18 14:29:47	Residence	Food
	2014-05-18 14:29:47	2014-05-18 14:42:10	Food	Food
	2014-05-18 14:42:10	2014-05-18 15:47:08	Food	Travel and Transport
	2014-05-18 15:47:08	2014-05-18 16:00:18	Travel and Transport	Travel and Transport
	2014-05-18 16:00:18	2014-05-18 16:32:40	Travel and Transport	Residence

Por fim, foram criadas matrizes de frequências de transições entre categorias, essas guardam informações a cerca do deslocamento diário de um usuário entre as categorias. Assim, é possível ter ideia do padrão de mobilidade de cada usuário dentro de sua cidade. No entanto, esse trabalho visa realizar um estudo em larga escala e para isso, foram sumarizadas as informações diárias de todos os usuários de uma cidade em uma só matriz que passa a representar a dinâmica social dessa cidade. Essa abordagem é semelhante a de [Silva et al. 2012b] que as utiliza para construir o que chama de *Image of City*. As tabelas 2, 3 e 4 representam um exemplo do processo aplicado a cidade de São Paulo nos dias 12/05/2014, 16/05/2014 e 17/05/2014, respectivamente.

Tabela 2. Segunda

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	13	3	6	1	1	6	4	7	1
C&U	1	28	18	0	6	13	29	10	7
FD	10	23	51	4	22	40	14	34	13
NL	0	2	2	1	1	1	2	1	0
O&R	4	6	20	0	19	24	19	14	18
P&O	7	21	67	2	17	53	29	26	18
RD	1	8	3	0	10	14	11	6	1
S&S	9	10	54	2	18	19	15	64	10
T&T	1	7	20	3	11	23	18	23	64

Tabela 3. Sexta

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	6	3	13	7	7	4	6	5	5
C&U	5	21	18	10	1	12	25	10	9
FD	16	14	72	31	11	35	21	45	10
NL	6	5	8	29	0	2	6	3	3
O&R	6	4	21	1	23	15	21	18	19
P&O	6	12	75	6	15	50	22	38	27
RD	6	5	13	6	9	8	17	7	9
S&S	15	6	81	6	16	18	17	59	11
T&T	8	10	29	10	18	38	14	10	68

Tabela 4. Sábado

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	8	2	26	14	12	4	2	3	5
C&U	3	2	15	0	4	5	3	9	3
FD	28	4	65	18	19	17	29	50	9
NL	6	0	7	12	3	0	2	1	1
O&R	13	5	24	7	27	7	16	23	11
P&O	5	2	29	4	7	19	17	20	8
RD	6	0	16	6	12	12	28	10	6
S&S	22	0	101	10	20	14	26	116	12
T&T	9	7	15	4	20	8	7	21	36

4. Comparações entre Matrizes de Frequência de uma Mesma Cidade

Como mostrado anteriormente, o processo de tratamento teve como resultado matrizes que representam o padrão de mobilidade das cidades. Essas matrizes são separadas por cidade e dia. Nessa seção, elas foram investigadas com o objetivo de responder a duas perguntas: (i) o comportamento social das pessoas independe do dia da semana? Ou seja, é possível agrupar as amostras diárias sem haver distorção da informação? É possível separar por dias de semana e finais de semana? (ii) A população segue uma rotina semanal? Ou melhor, é possível associar os dados de um mesmo dia da semana, mas de semanas diferentes?

Antes de responder tais perguntas, é preciso ter em mente que as cidades se comportam de maneiras diferentes em relação a cada categoria, ou seja, o comportamento social ao longo do tempo das pessoas quanto à alimentação (categoria *FD*), por exemplo, é logicamente diferente se comparado ao comportamento em relação ao trabalho (categoria *P&O*), o que pode ser facilmente percebido analisando verticalmente os histogramas da figura 2. Então, só é aconselhável considerar dois dias semelhantes se todas as categorias, de ambos os dias, tiverem comportamentos similares.

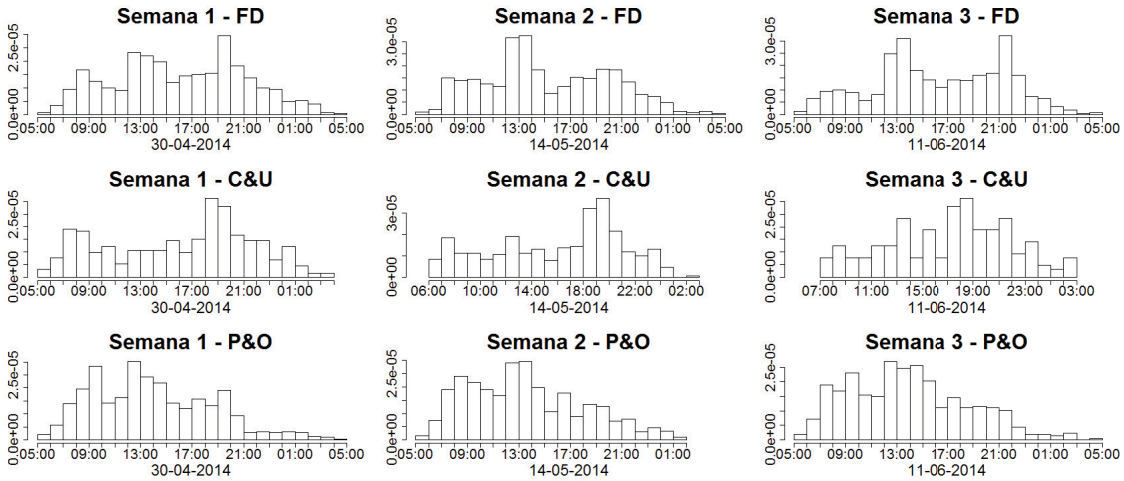

Figura 2. Histogramas da cidade de São Paulo apresentando o comportamento ao longo do tempo de 2 categorias em 3 diferentes semanas.

Dito isso, uma dedução lógica - e que pode ser facilmente confirmada verificando a diferença nos valores das tabelas 2, 3 ou 4 - é que, as chances de transições entre categorias são diferentes, isto é, a probabilidade dos usuários saírem de casa (categoria *RD*) para o trabalho (categoria *C&U*), por exemplo, é diferente da destes irem ao cinema (categoria *A&E*). No entanto, não é possível afirmar facilmente que tais probabilidades se mantêm independente do dia da semana.

Então, para responder a primeira pergunta, foi inicialmente selecionada uma cidade, para em seguida comparar, categoria-a-categoria e dia-a-dia, todas as matrizes de frequência em um período de uma semana. Para isso foi utilizado um teste Qui-quadrado, onde a hipótese nula pode ser enunciada como: as frequências de uma determinada linha l (categoria) da matriz M_{seg} são iguais as frequências da linha l da matriz M_{ter}. Esse teste foi executado para todos os dias da semana e para todas as cidades. Para analisar seu

resultado, foram criadas tabelas para cada categoria com o resultado do nível descritivo (*p-valor*) da comparação de todos os dias da semana.

Tabela 5. NY: Categoria *A&E*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	4.174e-02	2.007e-05	1.864e-05	8.785e-04	3.671e-08	1.166e-01
Terça	4.174e-02	1	1.650e-03	4.499e-02	3.622e-04	9.509e-07	3.808e-07
Quarta	2.007e-05	1.650e-03	1	4.351e-03	2.404e-02	1.640e-02	1.745e-03
Quinta	1.864e-05	4.499e-02	4.351e-03	1	7.654e-02	5.010e-11	3.468e-05
Sexta	8.785e-04	3.622e-04	2.404e-02	7.654e-02	1	2.735e-05	1.838e-06
Sábado	3.671e-08	9.509e-07	1.640e-02	5.010e-11	2.735e-05	1	1.408e-04
Domingo	1.166e-01	3.808e-07	1.745e-03	3.468e-05	1.838e-06	1.408e-04	1

Tabela 6. NY: Categoria *C&U*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	1.325e-01	1.592e-01	2.518e-02	8.763e-01	1.777e-01	8.973e-01
Terça	1.325e-01	1	3.968e-01	6.254e-01	8.302e-01	5.578e-01	2.636e-01
Quarta	1.592e-01	3.968e-01	1	5.749e-01	3.743e-01	7.847e-01	8.973e-01
Quinta	2.518e-02	6.254e-01	5.749e-01	1	4.470e-02	7.952e-01	2.433e-01
Sexta	8.763e-01	8.302e-01	3.743e-01	4.470e-02	1	4.232e-01	7.518e-01
Sábado	1.777e-01	5.578e-01	7.847e-01	7.952e-01	4.232e-01	1	7.518e-01
Domingo	8.973e-01	2.636e-01	8.973e-01	2.433e-01	7.518e-01	7.518e-01	1

Tabela 7. NY: Categoria *FD*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	7.972e-04	6.602e-13	6.601e-20	1.523e-18	5.714e-47	5.848e-11
Terça	7.972e-04	1	9.052e-04	3.381e-11	3.481e-07	2.415e-21	5.306e-05
Quarta	6.602e-13	9.052e-04	1	1.256e-13	2.842e-12	4.191e-35	3.331e-13
Quinta	6.601e-20	3.381e-11	1.256e-13	1	2.266e-07	1.820e-31	8.589e-11
Sexta	1.523e-18	3.481e-07	2.842e-12	2.266e-07	1	5.043e-04	1.669e-05
Sábado	5.714e-47	2.415e-21	4.191e-35	1.820e-31	5.043e-04	1	1.252e-08
Domingo	5.848e-11	5.306e-05	3.331e-13	8.589e-11	1.669e-05	1.252e-08	1

Tabela 8. NY: Categoria *NL*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	6.557e-01	7.122e-01	3.132e-01	1.623e-02	6.435e-08	2.659e-06
Terça	6.557e-01	1	7.180e-01	4.861e-01	3.620e-02	1.351e-05	5.078e-07
Quarta	7.122e-01	7.180e-01	1	2.616e-01	3.299e-02	3.254e-05	2.327e-05
Quinta	3.132e-01	4.861e-01	2.616e-01	1	5.802e-01	1.061e-07	4.072e-04
Sexta	1.623e-02	3.620e-02	3.299e-02	5.802e-01	1	7.615e-04	1.077e-01
Sábado	6.435e-08	1.351e-05	3.254e-05	1.061e-07	7.615e-04	1	2.155e-02
Domingo	2.659e-06	5.078e-07	2.327e-05	4.072e-04	1.077e-01	2.155e-02	1

Tabela 9. NY: Categoria *O&R*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	6.735e-02	7.499e-02	1.706e-02	1.471e-01	3.667e-06	1.766e-05
Terça	6.735e-02	1	6.769e-01	8.912e-02	5.507e-02	2.270e-05	7.171e-05
Quarta	7.499e-02	6.769e-01	1	2.887e-02	2.375e-02	9.105e-07	2.044e-07
Quinta	1.706e-02	8.912e-02	2.887e-02	1	5.958e-01	8.188e-03	1.452e-03
Sexta	1.471e-01	5.507e-02	2.375e-02	5.958e-01	1	4.180e-02	6.020e-03
Sábado	3.667e-06	2.270e-05	9.105e-07	8.188e-03	4.180e-02	1	2.487e-05
Domingo	1.766e-05	7.171e-05	2.044e-07	1.452e-03	6.020e-03	2.487e-05	1

Tabela 10. NY: Categoria *P&O*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	3.162e-01	2.814e-03	7.869e-02	4.240e-02	2.288e-05	2.269e-01
Terça	3.162e-01	1	3.001e-01	8.480e-01	9.817e-01	9.402e-01	1.528e-01
Quarta	2.814e-03	3.001e-01	1	7.929e-01	6.713e-01	3.349e-01	1.134e-01
Quinta	7.869e-02	8.480e-01	7.929e-01	1	5.464e-01	1.918e-01	7.875e-02
Sexta	4.240e-02	9.817e-01	6.713e-01	5.464e-01	1	6.327e-01	2.445e-01
Sábado	2.288e-05	9.402e-01	3.349e-01	1.918e-01	6.327e-01	1	1.733e-02
Domingo	2.269e-01	1.528e-01	1.134e-01	7.875e-02	2.445e-01	1.733e-02	1

Tabela 11. NY: Categoria *RD*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	4.258e-01	5.516e-01	5.874e-03	4.159e-01	6.747e-01	9.173e-01
Terça	4.258e-01	1	2.054e-01	3.204e-01	2.302e-02	5.825e-01	3.865e-01
Quarta	5.516e-01	2.054e-01	1	2.331e-12	1.539e-02	2.058e-01	8.140e-01
Quinta	5.874e-03	3.204e-01	2.331e-12	1	6.036e-02	9.651e-01	7.541e-01
Sexta	4.159e-01	2.302e-02	1.539e-02	6.036e-02	1	4.026e-01	7.358e-01
Sábado	6.747e-01	5.825e-01	2.058e-01	9.651e-01	4.026e-01	1	9.692e-01
Domingo	9.173e-01	3.865e-01	8.140e-01	7.541e-01	7.358e-01	9.692e-01	1

Tabela 12. NY: Categoria *S&S*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	7.866e-05	2.995e-04	1.085e-03	4.423e-11	1.077e-20	3.991e-10
Terça	7.866e-05	1	1.368e-01	4.765e-03	2.636e-04	2.969e-03	9.475e-03
Quarta	2.995e-04	1.368e-01	1	2.705e-02	2.137e-04	1.570e-02	1.819e-01
Quinta	1.085e-03	4.765e-03	2.705e-02	1	4.086e-04	1.717e-04	9.349e-02
Sexta	4.423e-11	2.636e-04	2.137e-04	4.086e-04	1	4.546e-03	9.416e-02
Sábado	1.077e-20	2.969e-03	1.570e-02	1.717e-04	4.546e-03	1	8.935e-03
Domingo	3.991e-10	9.475e-03	1.819e-01	9.349e-02	9.416e-02	8.935e-03	1

Tabela 13. NY: Categoria *T&T*

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Segunda	1	4.186e-02	1.623e-01	3.063e-02	2.014e-01	1.330e-04	1.343e-02
Terça	4.186e-02	1	1.502e-01	3.269e-02	1.294e-01	1.705e-05	6.813e-01
Quarta	1.623e-01	1.502e-01	1	5.260e-02	3.721e-02	1.065e-08	4.859e-03
Quinta	3.063e-02	3.269e-02	5.260e-02	1	4.503e-04	5.566e-08	5.337e-03
Sexta	2.014e-01	1.294e-01	3.721e-02	4.503e-04	1	2.745e-02	8.483e-03
Sábado	1.330e-04	1.705e-05	1.065e-08	5.566e-08	2.745e-02	1	1.678e-07
Domingo	1.343e-02	6.813e-01	4.859e-03	5.337e-03	8.483e-03	1.678e-07	1

As tabelas 5, 6, 7, 8, 9, 10, 11, 12 e 13, representam os valores do teste para a cidade de New York (NY). Verificando alguns resultados é possível perceber que, por exemplo, o comportamento da categoria *C&U* (Tabela 6), tende sempre a se manter bastante regular em todos os dias da semana, inclusive nos finais de semana, onde a população deve frequentar bibliotecas e outros locais de estudo. Esse comportamento não se repete em muitas cidades, por exemplo nas cidades brasileiras analisadas há uma clara diferença entre dias de semana e finais de semana.

Já a categoria *NL* (tabela 8), segue um padrão de comportamento bastante homogêneo durante os dias de semana, no entanto no final de semana há um grande aumento nos valores do teste. Esse resultado é bastante plausível, pois, durante a semana, por conta do trabalho, a população tende a não frequentar tanto a vida noturna, no entanto, essa atividade aumenta bastante em finais de semana, em especial sábado. Curiosamente, é possível perceber pela tabela, que o comportamento da população no domingo se assemelha muito mais com o de sexta, esse resultado provavelmente se deve por conta que, tanto o domingo quanto a sexta são dias que nem todos escolhem de sair a noite, já que muitos trabalham sábado e querem descansar no domingo.

Na categoria *S&S* (Tabela 12), é notado um alto grau de desordem entre os resultados, esse comportamento apesar de inconveniente, já era esperado, pois se trata de uma categoria relacionada a compras onde não existe uma periodicidade de comportamento, principalmente se tratando de uma cidade do tamanho de NY.

De fato, analisando todas as tabelas é possível perceber que o comportamento varia muito de acordo com a categoria estudada, sendo possível, normalmente, dividi-las em 3 grupos: (1) categorias *C&U*, *P&O* e *RD*, tem um comportamento mais linear, pois representam a rotina diária da população; (2) *S&S* e *FD*, tem uma maior variação, uma vez que são atividades que não fazem parte do dia-a-dia; (3) *A&E*, *NL*, *O&R* e *T&T* são atividades que seguem padrões bem diferentes em dias de semana e em finais de semana, já que representam, em geral, atividades de lazer.

Como foi possível ver, apenas algumas das categorias mantiveram um padrão ao longo dos dias, as demais têm uma dinâmica bastante diferente. Definindo um nível de significância para o teste de 5% por exemplo, no grupo 1, não existem evidências conclusivas, na grande maioria dos resultados, para rejeitar o teste. No entanto, no grupo 2 existem claras evidencias para se rejeitar o teste em praticamente todos os resultados. Já no grupo 3, as categorias aparentam poder serem associadas nos dias de semana, contudo, em finais de semana seguem um padrão de mobilidade bastante diferente. Em suma, conforme dito anteriormente, não é coveniente agregar dois dias a menos que todas as categorias destes sejam semelhantes, então nesse trabalho foram analisados todos os dias da semana separadamente.

Tabela 14. Los Angeles - S1 x S2

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
A&E	2.746e-01	2.767e-02	6.290e-02	8.488e-01	1.686e-02	6.284e-02	7.303e-02
C&U	9.460e-01	3.307e-01	2.564e-01	4.895e-01	6.149e-01		5.637e-01
FD	1.337e-02	7.026e-11	1.164e-02	5.326e-02	2.092e-04	2.243e-01	1.987e-01
NL	8.052e-01	3.509e-01	3.813e-01	6.873e-02	1.340e-05	5.106e-01	1.178e-04
O&R	9.263e-01	6.604e-02	1.112e-03	9.430e-02	2.668e-05	5.531e-01	2.172e-02
P&O	3.705e-02	5.248e-01	9.649e-01	4.444e-01	7.102e-01	5.886e-01	6.547e-01
RD	7.893e-01	6.948e-02	6.881e-01	7.745e-02	2.865e-01	8.614e-01	6.939e-01
S&S	4.326e-01	7.665e-04	7.653e-02	8.367e-07	1.723e-03	5.736e-03	7.513e-02
T&T	5.439e-02	4.869e-01	6.497e-01	2.383e-02	3.102e-04	2.033e-02	1.609e-01

Tabela 15. Los Angeles S2 x S3

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
A&E	9.665e-01	8.898e-03	6.426e-02	9.067e-01	5.600e-01	1.486e-02	3.486e-02
C&U	3.719e-02	8.825e-01	3.173e-01	5.153e-01	8.254e-01		
FD	2.716e-01	8.048e-04	5.207e-05	3.147e-03	2.157e-05	3.167e-01	2.675e-01
NL	7.220e-01	7.746e-02	7.766e-01	1.798e-05	1.166e-01	7.420e-05	1.311e-01
O&R	1.083e-01	1.110e-01	6.540e-05	1.668e-01	7.582e-01	1.144e-02	3.320e-02
P&O	5.878e-02	6.473e-01	1.857e-02	9.796e-06	7.003e-01	5.866e-01	7.237e-01
RD	7.912e-01	8.816e-02	8.326e-02	3.679e-01	8.486e-02	7.578e-01	7.710e-02
S&S	9.013e-01	2.032e-10	4.455e-02	3.967e-01	5.210e-01	2.006e-02	7.809e-02
T&T	7.304e-01	1.958e-04	1.869e-02	3.355e-02	3.522e-04	3.886e-01	4.099e-01

Em relação à dúvida quanto a população seguir ou não uma rotina semanal, foram separadas 3 semanas com amostra diárias de todas as categorias e, de maneira análoga à anterior, executou-se um teste Qui-quadrado para comparar , novamente categoria a categoria, o mesmo dia da semana, das 3 diferentes semanas selecionadas.

Analisando os resultados do exemplo da cidade de Los Angeles (tabelas 14 e 15), foi possível perceber que na comparação da semana *S1* com a semana *S2*, para nível de significância de 5%, não existem evidências para se refutar o teste em aproximadamente 68% dos casos, se aumentado o nível de significância para 1% esse valor sobe para 81%. Já na comparação das semanas *S2* e *S3* esse valores são de, 58% e 76% para níveis de significância de 5% e 1% respectivamente. Vale salientar que os valores em branco nas tabelas não atingiram as condições mínimas para uma execução fidedigna do teste.

Verificando os números acima é possível perceber que existe uma semelhança entre as semanas. As falhas no teste podem representar ruídos nas amostras coletadas, variabilidade intrínseca da população estudada e até mesmo um evento ocorrido em um dado dia classificado em uma categoria específica. Nesse trabalho foi realizado a junção das semanas, mas em trabalhos futuros pretende-se estudar a evolução dos dados ao longo

do tempo, no entanto para isso se faz necessário uma amostra com uma fração de tempo maior que a utilizada.

4.1. Comparações entre Matrizes de Frequência de Diferentes Cidades

Extrapolando a análise na seção anterior, foram comparadas, por categorias e dias da semana, diferentes cidades com o objetivo de se descobrir quais cidades se parecem em relação a características específicas.

A cidade de São Paulo foi escolhida como exemplo, a tabela 16 apresenta os resultados da execução do teste Qui-quadrado para comparação com as demais cidades e suas respectivas categorias, em um dia de segunda-feira, apenas para ilustrar o método.

Nesse caso os *p-valores* foram utilizados para rankear os resultados. Essa mesma análise foi realizada para todos os dias da semana, e foi feito um ranking das cidades que mais se assemelham à São Paulo em cada categoria, ou seja, que mais vezes aparecem como mais semelhantes durante os outros dias da semana ou do fim de semana. Alguns resultados desta análise merecem destaque:

- Levando em consideração todos os dias úteis, as cidades que mais próximas relação à categoria *P&O* - que representa todas as categorias ligadas atividades de trabalho - são: Moscow, Los Angeles e Mexico City.
- Ainda tendo em conta apenas dias úteis, a categorias ligadas a educação (*C&U*) as que mais parecidas são: Moscow, Rio de Janeiro e Recife.
- Já em relação a finais de semana, na categoria que classifica locais de vida noturna, como bares e casas de show (categoria *NL*), são: Bangkok, Rio de Janeiro e St. Petersburg.
- Também em relação a finais de semana, mas na categoria que representa os hábitos alimentares da populaçao (Categoria *FD*), as cidades mais parecidas são: Moscow, Recife e Rio de Janeiro.
- E por fim, em finais de semana, as cidades mais próximas na categoria que representa locais que oferecem conteúdo de arte e entretenimento, como museus, cinemas e teatros, são: Los Angeles, Moscow e Mexico City.

Observe que os resultados discutidos acima nem sempre se refletem na tabela 16, pois nesta análise também estão incluídos as tabelas dos outros dias da semana (as outras tabelas foram omitidas por limitação de espaço).

Tabela 16. São Paulo: Segunda-feira

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
Bangkok	4.128e-03	1.182e-20	1.276e-04	8.595e-04	2.505e-14	1.869e-04	1.234e-05	1.805e-16	2.051e-12
Los Angeles	8.016e-04	1.347e-04	5.219e-17	3.624e-02	2.909e-06	5.544e-06	9.377e-03	2.370e-15	1.572e-29
Manila	9.874e-02	6.360e-15	1.898e-05	8.316e-01	7.100e-08	2.981e-12	8.317e-08	1.358e-37	3.051e-11
Mexico City	3.056e-02	6.841e-11	2.084e-02	5.587e-01	4.583e-02	6.971e-10	7.105e-04	1.948e-01	4.341e-12
Moscow	3.655e-01	2.586e-01	1.400e-08	9.811e-01	1.518e-02	6.168e-06	7.067e-08	1.034e-04	3.874e-05
New York	5.860e-03	4.498e-23	1.215e-03	5.015e-02	1.353e-03	4.746e-08	2.465e-12	2.638e-05	1.628e-52
Recife	3.284e-01	2.577e-03	6.002e-10	4.724e-01	3.425e-13	6.615e-07	1.313e-01	6.673e-10	3.296e-07
Rio de Janeiro	3.181e-01	4.477e-02	2.172e-14	5.637e-01	7.860e-02	1.944e-09	1.120e-02	2.165e-07	6.766e-28
St. Petersburg	1.557e-04	1.218e-07	2.412e-15	3.458e-01	8.977e-03	6.903e-07	2.613e-05	6.459e-10	4.222e-07

4.2. Comparações entre Matrizes de Markov de Diferentes Cidades

Ao estudar o comportamento das cidades separando-as em categorias foi possível entender a permutação entre uma categoria específica e as demais, no entanto, para poder

realmente entender a dinâmica de uma cidade é necessário estudar o comportamento da população em relação a todas as categorias em conjunto.

Para retratar tal comportamento foram utilizadas *Cadeias de Markov em tempo discreto (DTMC)*, onde cada categoria passa a representar um estado das *Matrizes de Transição de Probabilidade*, como logicamente não é possível contar com toda a população necessária para construir tais matrizes, estas foram estimadas a partir dos dados amostrais coletados de cada cidade.

A escolha por *DTMC* foi feita em especial por três características encontradas nestes tipos cadeias, (i) é um processo estocástico bastante utilizado para representar estados discretos; (ii) tem capacidade de representar semelhantemente dados com frequências de amostragem diferentes; e (iii) em *Processos Markovianos* a probabilidade condicional de qualquer evento futuro depende somente do estado presente, sendo assim mais fácil de se representar dados temporais quando se tem apenas amostras de tempo.

Existem diferentes métodos para se estimar as *Matrizes de Transição* de *DTMCs*, neste trabalho utilizamos a abordagem descrita em [Pardo 2005], que apresenta um estimador de máxima verossimilança para as probabilidades de transição de estado da matriz de transição.

É importante ressaltar que o estimador descrito em [Pardo 2005] é um estimador de *Máxima Verossimilhança*, sendo este consistente, porém tendencioso, com o viés tendendo a zero a medida que o tamanho da amostra aumenta. Assim, é possível estimar uma *Matriz de Transição* de maneira consistente tendo uma amostra grande o suficiente.

As tabelas 17, 18, 19 e 20 representam Matrizes de Transição de Probabilidade estimadas a partir de dados de dias de semana da cidade de New York, como podemos perceber, existem probabilidades associadas à transição entre todos os estados da *Cadeia de Markov*, então é possível por exemplo estimar a probabilidade dos habitantes de New York que saem dos seus trabalhos para irem a um bar durante um dia de sexta-feira.

Tabela 17. New York: Segunda

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	0.07143	0.017860	0.4286	0.07143	0.07143	0.07143	0.03571	0.16070	0.07143
C&U	0.12500	0.156200	0.2500	0.00000	0.03125	0.18750	0.03125	0.12500	0.09375
FD	0.05469	0.042970	0.2500	0.07031	0.11720	0.14840	0.07031	0.14840	0.09766
NL	0.05128	0.000000	0.2308	0.46150	0.05128	0.02564	0.10260	0.02564	0.05128
O&R	0.00000	0.021740	0.2826	0.07609	0.11960	0.14130	0.05435	0.13040	0.17390
P&O	0.06536	0.039220	0.2745	0.03268	0.14380	0.20920	0.03922	0.11110	0.08497
RD	0.05556	0.000000	0.2778	0.00000	0.11110	0.16670	0.05556	0.22220	0.11110
S&S	0.03093	0.041240	0.2268	0.02577	0.20620	0.11340	0.04639	0.24740	0.06186
T&T	0.02439	0.006098	0.2134	0.02439	0.11590	0.11590	0.02439	0.10370	0.37200

Tabela 18. New York: Quarta

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	0.26320	0.000000	0.4342	0.11840	0.02632	0.02632	0.00000	0.05263	0.07895
C&U	0.04762	0.190500	0.1905	0.14290	0.14290	0.14290	0.00000	0.14290	0.00000
FD	0.17650	0.010700	0.2620	0.09626	0.03209	0.17650	0.03209	0.12830	0.08556
NL	0.20000	0.000000	0.3714	0.37140	0.00000	0.00000	0.02857	0.02857	0.00000
O&R	0.02381	0.023810	0.2381	0.02381	0.16670	0.16670	0.09524	0.09524	0.16670
P&O	0.01639	0.016390	0.3525	0.04098	0.07377	0.23770	0.04918	0.11480	0.09836
RD	0.03846	0.038460	0.1154	0.07692	0.15380	0.26920	0.03846	0.15380	0.11540
S&S	0.03289	0.013160	0.3816	0.02632	0.11180	0.06579	0.05921	0.25000	0.05921
T&T	0.03390	0.008475	0.2034	0.01695	0.10170	0.16100	0.06780	0.06780	0.33900

Tabela 19. New York: Sexta

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	0.18350	0.000000	0.2752	0.20180	0.07339	0.02752	0.05505	0.07339	0.11010
C&U	0.00000	0.105300	0.3158	0.05263	0.05263	0.15790	0.05263	0.21050	0.05263
FD	0.10630	0.013570	0.3529	0.16290	0.09050	0.05882	0.02941	0.12440	0.06109
NL	0.13640	0.006494	0.2987	0.44810	0.01948	0.01299	0.02597	0.01299	0.03896
O&R	0.10500	0.015000	0.3100	0.07500	0.18500	0.09000	0.04000	0.11000	0.07000
P&O	0.05797	0.021740	0.3261	0.07246	0.13040	0.15220	0.03623	0.13770	0.06522
RD	0.04167	0.000000	0.2500	0.12500	0.04167	0.08333	0.12500	0.33330	0.00000
S&S	0.03030	0.012990	0.3117	0.05628	0.10820	0.03896	0.04762	0.37660	0.01732
T&T	0.03289	0.006579	0.1974	0.07237	0.08553	0.08553	0.04605	0.10530	0.36840

Tabela 20. New York: Domingo

	A&E	C&U	FD	NL	O&R	P&O	RD	S&S	T&T
A&E	0.15790	0.008772	0.3772	0.07895	0.13160	0.03509	0.008772	0.15790	0.04386
C&U	0.00000	0.000000	0.6000	0.00000	0.00000	0.40000	0.000000	0.00000	0.00000
FD	0.10850	0.000000	0.3359	0.08786	0.11110	0.05943	0.049100	0.19120	0.05685
NL	0.03846	0.019230	0.4038	0.38460	0.03846	0.00000	0.038460	0.01923	0.05769
O&R	0.11660	0.018400	0.3006	0.05521	0.27610	0.05521	0.018400	0.09202	0.06748
P&O	0.07368	0.000000	0.4947	0.04211	0.09474	0.03158	0.042110	0.15790	0.06316
RD	0.04167	0.000000	0.2500	0.12500	0.12500	0.00000	0.000000	0.25000	0.20830
S&S	0.06038	0.007547	0.3208	0.03396	0.10940	0.04528	0.067920	0.32080	0.03396
T&T	0.10490	0.000000	0.1888	0.02797	0.07692	0.08392	0.041960	0.11890	0.35660

Após todas as Matrizes de probabilidade terem sido estimadas, foi utilizado o teste de hipótese para *Cadeias de Markov* de [Kullback et al. 1962] com intuito de calcular o grau de similaridade entre uma dada cidade e as demais para todos os dias da semana. Na tabela 21 é possível ver os resultados do *p-valor* de tal teste aplicado para comparar todas

as cidades selecionadas à New York. Como o agrupamento é feito por dia da semana, existe uma pequena variação na ordem de proximidade das cidades de acordo com o dia analisado.

Tal variação por exemplo pode ser vista verificando os dias de segunda-feira - onde as três cidades mais próxima de New York são Los Angeles, Mexico City e Manila - e quinta-feira onde apesar da primeira se manter a mesma, a segunda passa a ser São Paulo, com Mexico City como terceira e Manila apenas como a sexta mais próxima.

Tabela 21. New York

	Segunda	Terça	Quarta	Quinta	Sexta	Sábado	Domingo
Bangkok	1.097e-02	4.572e-03	1.293e-03	4.869e-04	2.116e-03	1.203e-03	6.271e-03
Los Angeles	4.444e-02	5.584e-02	1.741e-02	5.440e-02	1.658e-01	1.527e-01	2.280e-01
Manila	1.596e-02	4.678e-03	1.068e-01	4.326e-03	6.587e-03	1.845e-03	3.346e-03
Mexico City	1.752e-02	1.957e-02	3.363e-02	2.489e-02	6.835e-02	1.176e-01	5.227e-02
Moscow	1.518e-04	1.667e-03	8.301e-04	6.633e-03	2.435e-03	2.174e-03	1.132e-03
Recife	1.166e-04	6.562e-04	1.004e-05	6.396e-04	2.427e-05	1.343e-06	6.031e-05
Rio de Janeiro	7.598e-05	1.239e-04	5.212e-05	2.826e-03	7.868e-04	2.239e-05	1. 008e-06
Sao Paulo	6.612e-03	1.579e-03	1.552e-03	3.414e-02	2.457e-02	3.123e-02	1.600e-02
St. Petersburg	1.339e-03	2.665e-04	1.267e-04	1.169e-02	5.100e-03	5.277e-03	1.383e-03

Ao término do experimento, serão brevemente discutidas na próxima seção, algumas aplicações que possam fazer uso de tais resultados.

5. Aplicação

Inúmeras aplicações podem se beneficiar do agrupamento de cidades. Dependendo da aplicação, pode ser interessante investigar o grau de semelhança entre as cidades. Por exemplo, a utilização da técnica de análise proposta pode ser utilizada como um engenho para sistemas de recomendação voltados ao turismo. Nesse caso, o estudo pode ser de duas formas diferentes: agrupando apenas algumas categorias ou com todas as categorias juntas.

Por exemplo, dado que um usuário já tenha viajado para Recife e gostado da cidade por conta de seus museus, praças históricas e locais relacionados a cultura, é possível então gerar um agrupamento de cidades parecidas com *Recife* nestes aspectos. Para isso precisa-se primeiro verificar quais dias da semana o usuário pretende passar no local e então realizar um processo semelhante ao feito para cidade de São Paulo na seção 4.1, utilizando os *p-valores* resultante do teste Qui-quadrado para as categorias *Arts & Entertainment e Outdoors & Recreation* e então agrupar o resultado utilizando como parâmetro os valores obtidos, como realizado no dendrograma da figura 3.

No entanto, se o usuário deseja examinar se todas as características das cidades são semelhantes as de Recife, para então decidir para qual cidade deseja visitar, basta, após selecionar os dias da viagem, fazer uma comparação entre Matrizes de markov, como a feita para cidade de New York na seção 4.2. No segundo agrupamento da figura 4 é possível ver essa abordagem aplicada à cidade de Recife nos dias de sexta e sábado.

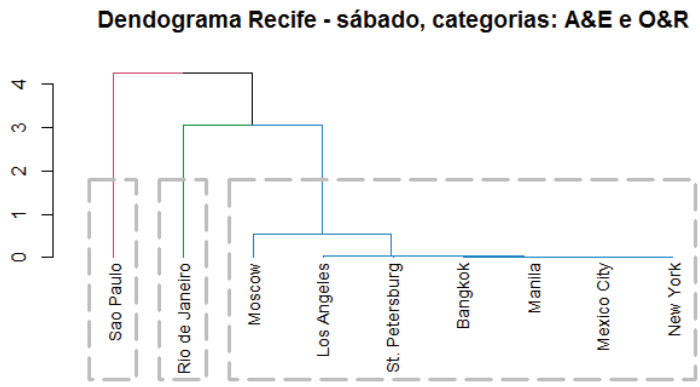

Figura 3. Agrupamento Hierárquico (método Centróide), exemplo de divisão em 3 grupos, usando *p-valor* do teste Qui-quadrado como parâmetro de agrupamento.

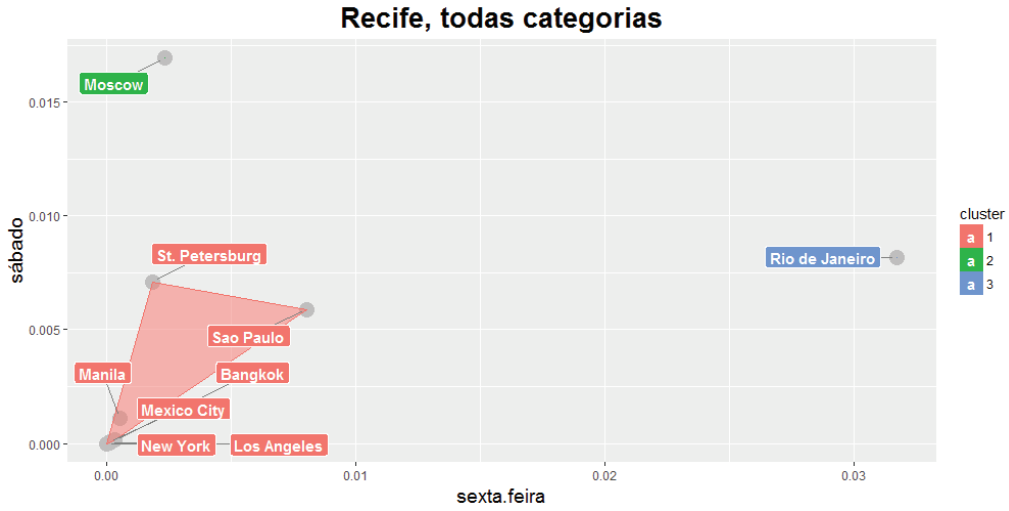

Figura 4. Agrupamento hierárquico (método Centróide), exemplo de divisão em 3 grupos, usando o *p-valor* do teste para Cadeias de Markov como parâmetro de agrupamento.

6. Conclusão

As técnicas de análise apresentadas podem servir para inúmeras aplicações, como estudo para viabilização de empreendimentos, cálculo de índice cultural da população, estudo de mercado e até análise de deslocamentos. A técnica de estudo populacional apresentada, que leva em consideração a dinâmica social da população nas cidades, é de fácil aplicação e, mesmo quando aplicada em larga escala, pode obter resultados em um curto espaço de tempo, ao contrário das técnicas tradicionais.

Referências

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. *International AAAI Conference on Weblogs and Social Media*, page 58.

Foursquare (2016). Foursquare website. Acessado em: 18-12-2016.

Foursquare (2017). Foursquare para desenvolvedores. Acessado em: 07-04-2017.

Frias-Martinez, V., Soto, V., Hohwald, H., and Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 239–248.

ICT Data and Statistics Division (2016). ICT Facts & Figures. *14th World Telecommunication/ICT Indicators Symposium (WTIS)*.

Inglehart, R. and Welzel, C. (2010). Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, 8(02):551–567.

Kosmin, B. A. and Lachman, S. P. (2011). *One Nation Under God: Religion in Contemporary American Society*. Crown Publisher.

Kostakos, V., Nicolai, T., Yoneki, E., O'Neill, E., Kenn, H., and Crowcroft, J. (2009). Understanding and measuring the urban pervasive infrastructure. *Personal and Ubiquitous Computing*, 13(5):355–364.

Kullback, S., Kupperman, M., and Ku, H. H. (1962). Tests for contingency tables and markov chains. *Technometrics*, 4(4):573–608.

Noulas, A., Scellato, S., Mascolo, C., and M., P. (2011). An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11:70–573.

Pardo, L. (2005). *Statistical inference based on divergence measures*. CRC Press.

Phithakkitnukoon, S. and Olivier, P. (2011). Sensing urban social geography using online social networking data. In *The Social Mobile Web*, pages 36–39.

Silva, T. H. (2014). *Large Scale Study of City Dynamics and Urban Social Behavior Using Participatory Sensor Networks*. PhD thesis, Universidade Federal de Minas Gerais (UFMG).

Silva, T. H., d. Melo, P. O. S. V., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2012a). Visualizing the invisible image of cities. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on*, pages 382–389.

Silva, T. H., de Melo, P. O., Almeida, J., Musolesi, M., and Loureiro, A. (2014). You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. *arXiv preprint arXiv:1404.1009*.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012b). Visualizing the Invisible Image of Cities. In *Proc. IEEE International Conference on Cyber, Physical and Social Computing*, Besancon, France.

Stewart, J. B. (2016). Facebook has 50 minutes of your time each day. it wants more. New York Times Online, acessado em: 04-03-2017.

Tammaru, T., Ham, M. V., Marcinczak, S., and Musterd, S. (2016). *Socio-Economic Segregation in European Capital Cities: East Meets West*. Routledge.

Waters, M. C. (1994). Ethnic and racial identities of second-generation black immigrants in new york city. *The International Migration Review*, 28(4):795–820.

Xie, X.-F. and Wang, Z.-J. Combining physical and participatory sensing in urban mobility networks.

Zheng, Y. and Zhou, X. (2011). *Computing with Spatial Trajectories*. Springer.