



FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

WAGNER AUGUSTO LOPES DE VARGAS

DATA SCIENCE & SEGURANÇA PÚBLICA:
padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São
Paulo.

SÃO PAULO
2019

WAGNER AUGUSTO LOPES DE VARGAS

DATA SCIENCE & SEGURANÇA PÚBLICA:

padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São Paulo.

Artigo apresentado à Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas, como um dos requisitos para obtenção do título de Mestre em Gestão e Políticas Públicas.

Campo de conhecimento: Segurança Pública; Análise de Dados; Estatística; Ciência de Dados.

Orientador: Prof. Dr. Nelson Marconi

SÃO PAULO

2019

“O conhecimento é o processo de acumular dados;
a sabedoria reside na sua simplificação”.

Martin H. Fischer

RESUMO

O presente artigo analisa 1.001.006— um milhão, mil e seis— Boletins de Ocorrências (B.O.), do Registro Digital de Ocorrências (R.D.O.), acerca de roubo de celular na cidade de São Paulo no período de 10 anos, desde o primeiro dia do mês de janeiro do ano de 2010 e até o último dia do mês de dezembro de 2018, a partir da base de dados da secretaria da Segurança Pública do Governo do Estado de São Paulo. A análise é exploratória e busca encontrar padrões estatísticos nas ocorrências de roubo de celulares na cidade e entender as chances e possíveis relações entre os casos e regiões em que ocorreu flagrante delito.

Palavras-chave: Inteligência Artificial. Aprendizado de Máquina. Políticas Públicas. Segurança Pública. Ciência de dados. Econometria Espacial.

ABSTRACT

This article analyzes 1.001.006 - one million, one thousand and six - Bulletins of Mobile Phone Theft (BO) at São Paulo's city from the first day of January 2010 until the last day of December 2018. The study was made with the Public Security dataset collected by Secretariat of the São Paulo State Government. This research analysis seeks to find statistical patterns on the occurrences of cell phone theft and to understand the chances and possible relationships between cases and regions in which a flagrant offense occurred.

Key-words: Artificial Intelligence. Machine Learning. Public Policy. Public Security. Data Science. Spatial Econometrics.

SUMÁRIO

PARTE I - MODELOS LINEARES	6
1 INTRODUÇÃO.....	6
2 PROPOSTA DE TRABALHO	7
2.1 Objetivo geral.....	7
2.2 Metodologia	7
3 OCORRÊNCIAS POR DISTRITO POLICIAL.....	9
3.1 Geral	9
3.2 Distritos por região	10
3.3 Ocorrências por Região ao longo do tempo	13
3.4 Estatísticas descritivas	14
3.5 Regressão Linear.....	21
PARTE II- MODELOS ESPACIAIS	28
3.6 Geral e descritivas	28
3.6.1 Matriz de vizinhança	29
3.6.2 Autocorrelação espacial.....	30
3.6.3 Regressão Espacial Global	33
PARTE III- MACHINE LEARNING.....	38
4 CHANCES DE FLAGRANTE	38
4.1 Geral	38
4.1.1 Treino e Teste	38
4.2 Regressão Logística.....	39
4.3 Árvore de Decisão	42
3.5 Bootstrap aggregating (Bagging).....	44
3.6 Random Forest	45
3.7 Support Vector Machine.....	46
3.8 Redes Neurais.....	47
5 CONCLUSÕES.....	49
REFERÊNCIAS.....	51

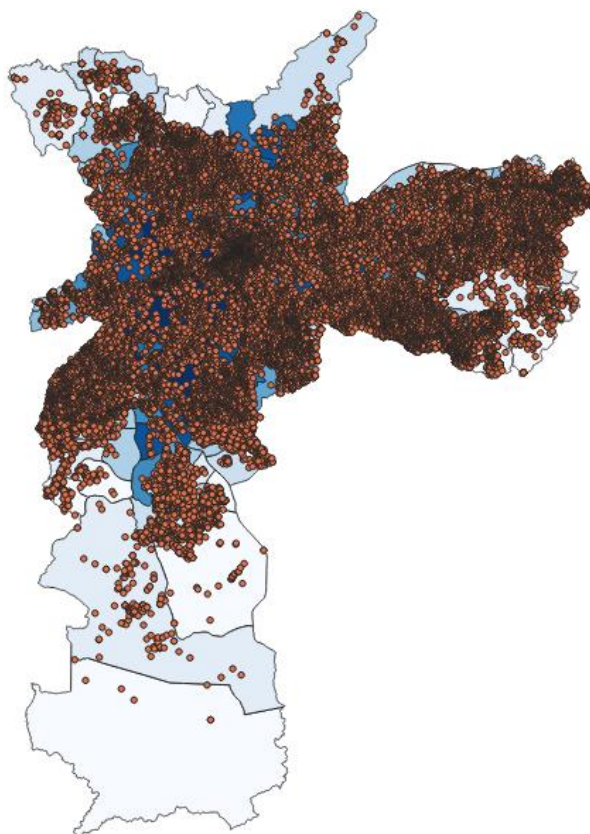
PARTE I - MODELOS LINEARES

1 INTRODUÇÃO

De acordo com Artigo 157 do Código Penal (Decreto Lei 2.848/1940), o roubo é um crime contra o patrimônio no qual ocorre a subtração de um bem móvel que também acontece violência ou ameaça grave de violência, de modo que a vítima não tenha condições de reagir:

Art. 157 – Subtrair coisa móvel alheia, para si ou para outrem, mediante grave ameaça ou violência a pessoa, ou depois de havê-la, por qualquer meio, reduzido à impossibilidade de resistência: Pena – reclusão, de quatro a dez anos, e multa.” (Decreto Lei 2.848/1940).

As ocorrências de roubo de aparelhos celular na cidade de São Paulo registradas pela Secretaria de Segurança Pública do Governo do Estado de São Paulo foram analisadas por este artigo à luz de ferramentas de Data Science.



Mapa 1– Locais onde ocorreram roubos de celular na cidade de São Paulo em 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2019)

2 PROPOSTA DE TRABALHO

2.1 Objetivo geral

Trata-se de um artigo cujo objeto central é realizar testes estatísticos para observar padrões e prever situações relacionadas referentes ao flagrante delito e ao total de ocorrências ou registros de roubos de celular na cidade de São Paulo. O objetivo é que alguns dos insights gerados possam contribuir para novos debates para implementar melhorias nas políticas públicas relativas ao tema em questão.

2.2 Metodologia

Além da análise das estatísticas descritivas, foram utilizadas técnicas de análise divididas em três âmbitos:

I- Enfoque inferencial: com o objetivo de encontrar relações causais entre as variáveis X e a variável Y foram utilizadas Regressões Lineares múltiplas stepwise backward, Análise de variância, testes T e F, teste de normalidade, de correlação, teste de auto-correlação entre resíduos, de multicolinearidade e de heterocedasticidade com uma equação linear capaz de explicar a relação entre as preditoras e as preditas;

II- Enfoque preditivo: construídos algoritmos com o objetivo de prever determinadas situações descritas nos Boletins (como probabilidade de flagrante) foram utilizadas técnicas de Árvore de Decisão, Regressão Logística, Regressões Lineares, Random Forest, Bagging, Support Vector Machines e Redes Neurais. O objetivo da análise é buscar melhor desempenho preditivo para prever ocorrências de flagrante delito;

III: Espacial: também com enfoque inferencial com objetivo de reduzir a auto-correlação espacial entre os resíduos, aumentar o R^2 e diminuir o Critério de Informação de AKAIKE, a fim de encontrar uma equação espacial que explique a variação de Y. Também foram realizadas análises de estatística espacial descritiva. Para essa seção foram utilizadas Spatial Lag e Spatial Error.

Para tais análises foram utilizados softwares, R Studio, Jupyter com a linguagem Python 3, Minitab, SPSS, Knime, Excel, Qgis e GeoDa. Também foram utilizadas técnicas webscrapping na base de dados da Secretaria da Segurança Pública do Estado de São Paulo.

A análise é aplicada único-exclusivamente nas bases de dados do Registro Digital de Ocorrências (R.D.O.), da Polícia Civil, disponibilizadas pela Secretaria de Segurança Pública em seu portal da transparência. De modo que não levados em consideração os dados do Sistema Interno de Ocorrências da Polícia Militar (SIOPM), que compete com o RDO, ainda que venha ocorrendo esforço por parte do Estado em promover integração entre ambos. A Escolha pelo RDO se dá pela disponibilidade pública deste Dataset. Há de ressaltar ainda que o Registro Digital de Ocorrência em tela analisado não leva em conta as ocorrências registradas por meio do telefone, através do 190.

3 OCORRÊNCIAS POR DISTRITO POLICIAL

3.1 Geral

Dentre o período analisado, que vai de janeiro de 2010 a dezembro de 2018, totalizaram-se 1.001.006— um milhão mil e seis— ocorrências de roubo de celular registradas nos 93 Distritos Policiais da cidade.

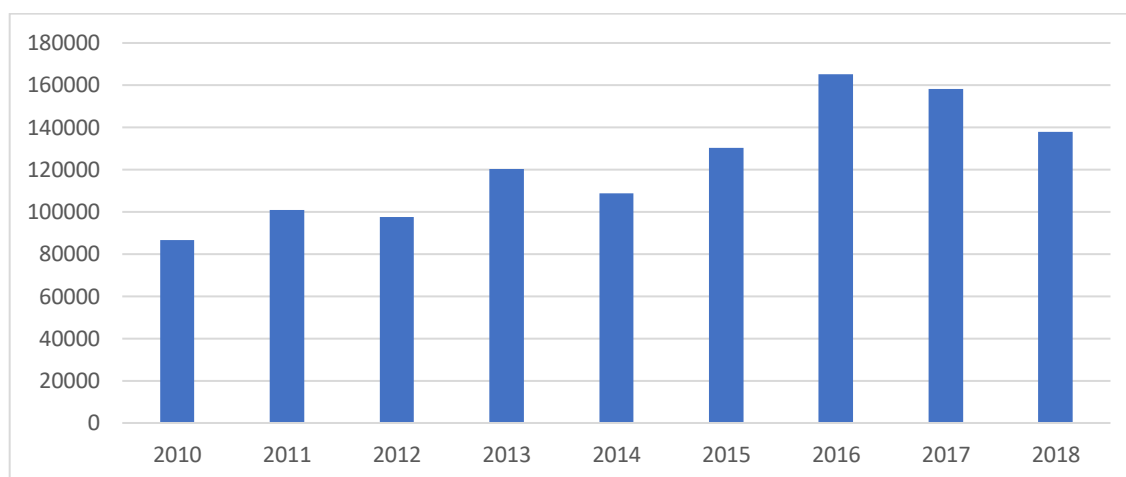


Gráfico 1– Evolução do Total de ocorrências policiais de roubos de celular na cidade de São Paulo entre 2010 e 2018 nos Distritos Policiais.

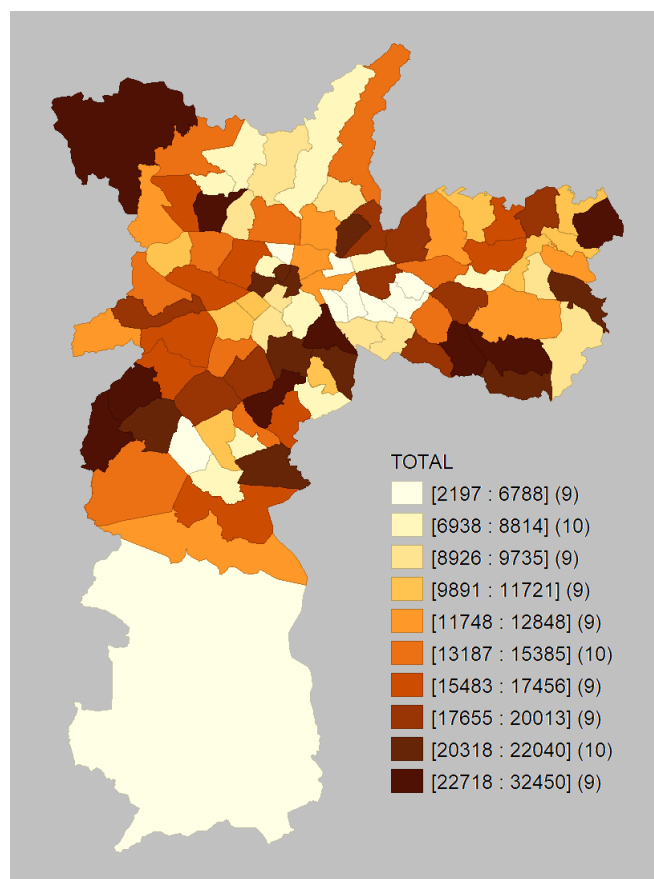
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

No decil com mais ocorrências de roubo de celular registradas na cidade no período, dentre os 9 Distritos Policiais, 4 deles estão na Zona Sul; 3 na Zona Leste e 2 na Zona Norte da Capital.

Tabela 1– decil de delegacias maior quantidade total de ocorrências de roubo de celular, de 2010 a 2018.

Rank	DELEGACIA	Zona	Ocorrências
1º	17º D.P. DOUTOR ALDO GALIANO/ Ipiranga	Sul	32450
2º	35º D.P. JABAQUARA	Sul	26836
3º	69º D.P. TEOTONIO VILELA	Leste	26183
4º	37º D.P. CAMPO LIMPO	Sul	25824
5º	49º D.P. SAO MATEUS	Leste	25778
6º	50º D.P. ITAIM PAULISTA	Leste	25255

7º	46º D.P. PERUS	Norte	24106
8º	47º D.P. CAPAO REDONDO	Sul	22982
9º	28º D.P. FREGUESIA DO O	Norte	22718



Mapa 2– Totalidade de ocorrências de roubo de celular na cidade de São Paulo por Distrito Policial, desde 2010 a 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.2 Distritos por região

A distribuição dos 93 Distritos Policiais (DP) dentre as Zonas da Capital demonstra que a Zona Leste é a que detém o maior número de Distritos Policiais, mas é também a zona que possui o maior número habitantes, logo a maior proporção de DP por morador.

Em termos de proporção a Zona Central tem o menor número de habitantes para serem atendidos no entorno de cada distrito, e é a região com o menor número de DP.

Tabela 2– Distritos Policiais por habitante e por região da cidade

<i>Zona</i>	Habitantes (IBGE)	Distritos Policiais	Habitante/Distrito
<i>Leste</i>	3.772.402	33	114
<i>Sul</i>	3.211.818	27	118
<i>Norte</i>	2.147.972	16	134
<i>Oeste</i>	920.806	9	102
<i>Centro</i>	373.914	8	47

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Tabela 3– Distritos Policiais por região da cidade de São Paulo.

Leste	33
08° D.P. BRAS	
103° D.P. COHABII/ITAQU.	
10° D.P. PENHA DE FRANCA	
18° D.P. ALTO DA MOOCA	
21° D.P. VILA MATILDE	
22° D.P. SAO MIGUEL PTA	
24° D.P. PONTE RASA	
29° D.P. VILA PRUDENTE	
30° D.P. TATUAPE	
31° D.P. VILA CARRAO	
32° D.P. ITAQUERA	
41° D.P. VILA RICA	
42° D.P. PARQUE S. LUCAS	
44° D.P. GUAIANAZES	
49° D.P. SAO MATEUS	
50° D.P. ITAIM PAULISTA	
52° D.P. PARQUE S.JORGE	
53° D.P. PARQUE DO CARMO	
54° D.P. CID. TIRADENTES	
56° D.P. VILA ALPINA	
57° D.P. PARQUE DA MOOCA	
58° D.P. VILA FORMOSA	
59° D.P. JARDIM DOS IPES	
62° D.P. ERMELINO MATARAZZO	
63° D.P. VILA JACUI	
64° D.P. CID.AE CARVALHO	
65° D.P. ARTUR ALVIM	
66° D.P. JDIM ARICANDUVA	
67° D.P. JARDIM ROBRU	
68° D.P. LAGEADO	
69° D.P. TEOTONIO VILELA	
70° D.P. VILA EMA	
81° D.P. BELEM	
Sul	27
100° D.P. JDIM HERCULANO	
101° D.P. JDIM IMBUIAS	

102° D.P. SOCORRO	
11° D.P. SANTO AMARO	
16° D.P. VILA CLEMENTINO	
17° D.P. DOUTOR ALDO GALIANO	
25° D.P. PARELHEIROS	
26° D.P. SACOMA	
27° D.P. IBIRAPUERA	
34° D.P. MORUMBI	
35° D.P. JABAQUARA	
36° D.P. VILA MARIANA	
37° D.P. CAMPO LIMPO	
43° D.P. CIDADE ADEMAR	
47° D.P. CAPAO REDONDO	
48° D.P. CIDADE DUTRA	
55° D.P. PARQUE S. RAFAEL	
80° D.P. VILA JOANIZA	
83° D.P. PARQUE BRISTOL	
85° D.P. JARDIM MIRNA	
89° D.P. JARDIM TABOAO	
92° D.P. P. STO ANTONIO	
95° D.P. HELIÓPOLIS	
96° D.P. MONÇÕES	
97° D.P. AMERICANOPOLIS	
98° D.P. JARDIM MIRIAM	
99° D.P. CAMPO GRANDE	
Norte	16
09° D.P. - CARANDIRU	
13° D.P. CASA VERDE	
19° D.P. VILA MARIA	
20° D.P. AGUA FRIA	
28° D.P. FREGUESIA DO O	
33° D.P. PIRITUBA	
38° D.P. VILA AMALIA	
39° D.P. VILA GUSTAVO	
40° D.P. VILA STA MARIA	
45° D.P. V BRASILANDIA	
46° D.P. PERUS	
72° D.P. VILA PENTEADO	
73° D.P. JACANA	
74° D.P. PARADA TAIPAS	
87° D.P. V. P. BARRETO	
90° D.P. PQ. NOVO MUNDO	
Oeste	9
07° D.P. LAPA	
14° D.P. PINHEIROS	
15° D.P. DR. LUC. H BEIGUELMAN	
23° D.P. PERDIZES	
51° D.P. RIO PEQUENO	
75° D.P. JARDIM ARPOADOR	
78° D.P. JARDINS	

91º D.P. CEASA	
93º D.P. JAGUARÉ	
Centro	8
01º D.P. SE	
02º D.P. BOM RETIRO	
03º D.P. CAMPOS ELISEOS	
04º D.P. CONSOLAÇÃO	
05º D.P. ACLIMACAO	
06º D.P. CAMBUCI	
12º D.P. PARI	
77º D.P. SANTA CECILIA	
Total Geral	93

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2019)

3.3 Ocorrências por Região ao longo do tempo

Observa-se que as zonas mais populosas apresentaram número de ocorrências crescente no decorrer dos anos. No entanto, em todas as cinco regiões, há uma queda de ocorrência no número de registros, ou a partir de 2016 ou de 2017. Movimento semelhante demonstrado no número de ocorrências total— que considera a soma de todas elas juntas. **Isso pode ser reflexo de uma diminuição no número de aparelhos roubados, mas também pode ser um indício de subnotificação de casos de roubo de celular.**

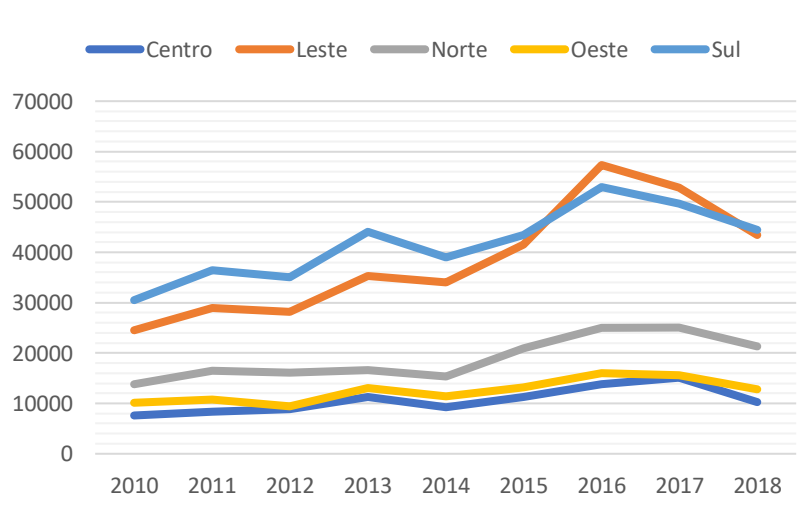


Gráfico 2– Evolução do Total de Ocorrências de roubos de celular na cidade de São Paulo entre 2010 e 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2019)

Há de destacar, também, possíveis impactos de políticas públicas implementadas pela secretaria da Segurança Pública do Estado de São Paulo nesta hipótese de queda das ocorrências. Desde 06/02/2015 a Polícia Civil registra o International Mobile Equipment Identity (IMEI) nos boletins de ocorrência dos celulares furtados e roubados. E, desde fevereiro de 2016, a Polícia Militar, de acordo com a Resolução SSP-3, passou a utilizar o código de IMEI como forma de fiscalização, realizando sua consulta por meio dos terminais móveis das viaturas. Ao ser registrado furto ou roubo, esse código, de acordo com a secretaria, também é enviado às operadoras de celulares que têm até 72h para realizar o bloqueio do aparelho.

A secretaria de Segurança Pública acredita que aumento de 209% de aparelhos recuperados, entre janeiro e outubro de 2016, em receptação sobre o mesmo período de 2015, deve-se também à coleta e utilização do IMEI nas ocorrências policiais.

Tabela 3– Evolução do Total de roubos de celular na cidade de São Paulo por região.

ANO	Centro	Leste	Norte	Oeste	Sul
2010	7600	24517	13806	10103	30498
2011	8333	28965	16512	10725	36376
2012	8860	28120	16080	9419	35025
2013	11303	35338	16570	13065	44083
2014	9263	34073	15277	11340	38924
2015	11314	41464	20943	13144	43392
2016	13835	57322	25050	16003	52946
2017	15106	52777	25055	15655	49689
2018	3731	10390	5126	3410	10473

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.4 Estatísticas descritivas

O total de ocorrências não apresenta outliers, nem mesmo a partir do recorte das cinco regiões geográficas da cidade, conforme indica o boxplot abaixo.

No entanto, pode enxergar-se maior dispersão e limite interquartilico no número de ocorrências registrado pelos Distritos Policiais da Zona Leste e da Zona Sul, entre 2010 e 2019. E a mediana de ocorrências dos DP da Zona Sul é, no mínimo, duas vezes maior do que a mediana de registros dos Distritos Policiais das Zonas Norte, Oeste e Centro.

Quanto ao número total de ocorrências em cada Distrito Policial, nos últimos 10 anos, também se observa que não há outliers, sendo que o número total de ocorrências de roubo de

celular nos 23 distritos figura no quartil superior. Deste modo, os cinco primeiros deles são das Zona Sul e Zona Leste, conforme lista abaixo em ordem decrescente de número de Boletins de Ocorrência, de roubos de celular:

17º D.P. Doutor Aldo Galiano; 69º D.P. Teotônio vilela; 37º D.P. Campo limpo; 49º D.P. São Mateus; 50º D.P. Itaim paulista; 46º D.P. Perus; 47º D.P. Capão redondo; 28º D.P. Freguesia do o; 03º D.P. Campos Elíseos; 04º D.P. Consolação; 19º D.P. Vila maria; 16º d.p. Vila clementino; 01º D.P. Sé; 44º D.P. Guaianazes; 95º D.P. Heliópolis; 98º D.P. Jardim Miriam; 55º D.P. Parque São Rafael; 30º D.P. Tatuapé; 11º D.P. Santo amaro; 51º D.P. Rio Pequeno; 22º D.P. São Miguel Paulista; 101º D.P. Jardim Imbuias; 34º D.P. Morumbi.

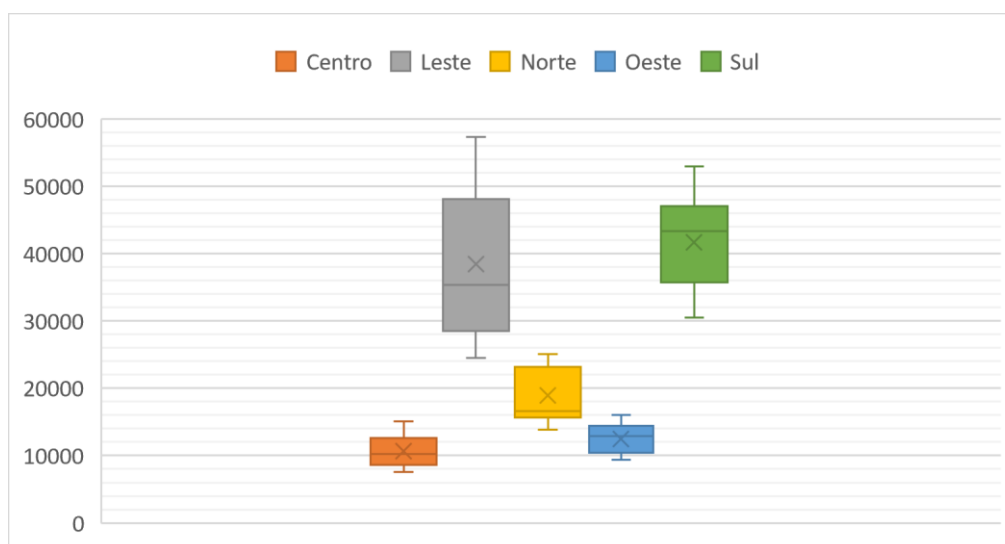


Gráfico 3– Boxplot Total de ocorrências de roubo de celular por ano em cada região da cidade de São Paulo

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Dentre as 33 variáveis analisadas, extraídas dos dados fornecidos em cada um dos Boletins de Ocorrência analisados, a variável “TOTAL” —que representa o total de ocorrências de roubos de celular registradas pela Secretaria de Segurança Pública do Estado de São Paulo — possui correlação forte com 13 outras variáveis, conforme indica tabela abaixo, cujo coeficiente é maior do que 0.7 ou 70%, sendo ele negativo ou positivo.

Em todas essas 13 variáveis— descritas na tabela com coeficiente decrescente— a Correlação de Pearson (dada pela equação abaixo) é significativa de forma bilateral e não ocorrem ao acaso, pois a probabilidade dessas variações ocorrerem está em um nível de confiança bilateral de 99%.

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}}$$

Ainda que não ocorra ao acaso, esse indicador não analisa relação de causalidade entre as variáveis, mas sim apenas demonstra que as duas respectivas variáveis analisadas variam simultaneamente, no mesmo sentido ou em sentidos opostos.

Tabela 4– Matriz de correlação entre total de ocorrências de Roubo de Celular e as demais variáveis.

	<i>TOTAL</i>		
	Correlação de Pearson	Sig. (bilateral)	N
<i>CONSUMADO</i>	.999**	.000	93
<i>DESCONHECIDA</i>	.992**	.000	93
<i>PUBLICA</i>	.932**	.000	93
<i>NOITE</i>	.877**	.000	93
<i>NOKIA</i>	.838**	.000	93
<i>DESCONHECIDO</i>	.832**	.000	93
<i>SAMSUNG</i>	.825**	.000	93
<i>MOTO</i>	.823**	.000	93
<i>SONY</i>	.804**	.000	93
<i>MANHA</i>	.800**	.000	93
<i>OUTRAS_M</i>	.791**	.000	93
<i>TARDE</i>	.758**	.000	93
<i>OUTROS_L</i>	.741**	.000	93
<i>MADRUGADA</i>	.552**	.000	93
<i>CONHECIDA</i>	.495**	.000	93
<i>COMERCIO</i>	.458**	.000	93
<i>FLAGRANTE</i>	.440**	.000	93
<i>APPLE</i>	.427**	.000	93
<i>BB</i>	.345**	.001	93
<i>INCERTA</i>	.344**	.001	93
<i>TENTADO</i>	.278**	.007	93
<i>RESTAURANTE</i>	.261*	.012	93
<i>FAVELA</i>	.223*	.032	93
<i>TOTAL</i>	1		93
<i>ESTRADA</i>	.151	.150	93
<i>BANCO</i>	.148	.156	93
<i>ESCOLA</i>	.131	.212	93
<i>SHOPPING</i>	.124	.236	93
<i>ESCRI</i>	.110	.293	93

<i>C_COMERCIAL</i>	.050	.631	93
<i>CONDOMINIO</i>	.035	.740	93
<i>REP_PUBLICA</i>	.030	.775	93
<i>INDUSTRIA</i>	-.073	.485	93

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

O Coeficiente de correlação de Pearson entre “TOTAL” e “CONSUMADO” — que mensura o número de roubos de celular em que o bandido obteve êxito na empreitada— é alto e positivo, registrado em 0,999 ou de 99,9%. Enquanto que a correlação entre “TOTAL” e “TENTADO” — a tentativa infrutífera de roubo do aparelho celular — tem coeficiente muito baixo, de 27,8%.

Abre-se, então, uma possibilidade para o questionamento: se o crime foi tentado sem êxito pode ter sido um incentivo negativo para a vítima registrar a ocorrência. O ponto é que, além de servir de indicador, a quantidade de B.O. pode ter alguma influência para determinar o patrulhamento policial da região; deste modo, se o crime tentado sem sucesso tende a ser menos registrado, pode perder-se uma oportunidade para prevenção de novos roubos de aparelhos celulares.

Entrementes, ainda que ele tenha relação forte e essa relação seja linear, o total de crimes consumado não pode ser considerado um bom preditor ou uma regressão Linear simples com a variável X “CONSUMADO” não é confiável para prever o total de ocorrências, uma vez que o resultado do teste de Breusch-Pagan é 0.000. Portanto, não refuta a hipótese de heterocedasticidade, de modo que a variância dos resíduos não é constante e há resíduos que são Upper outliers, conforme indica o gráfico de resíduos normalizados abaixo, com pontos plotados acima de 3.

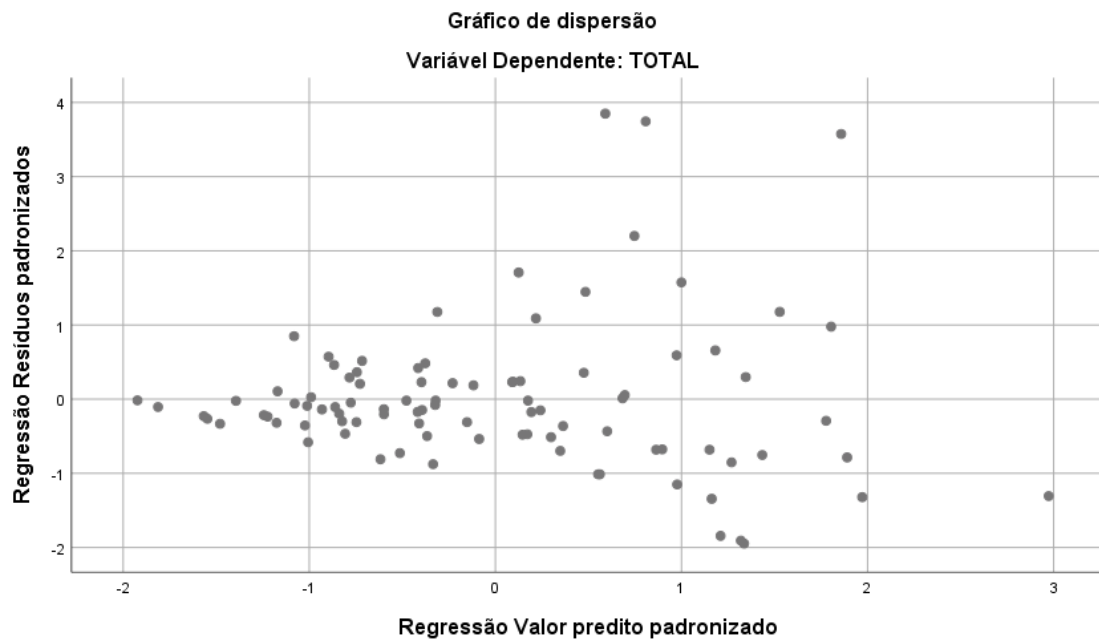
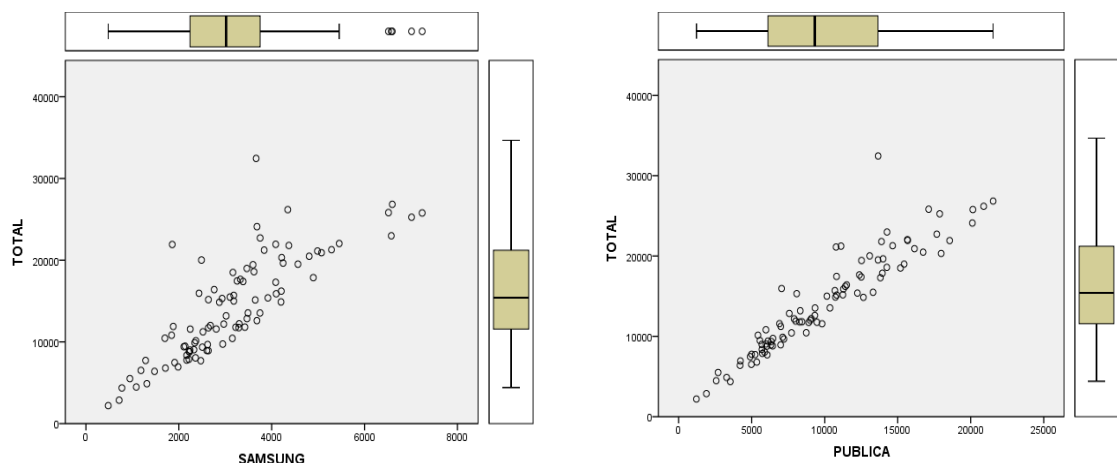


Gráfico 4— Resíduos padronizados contra valores preditos padronizados de regressão linear simples entre BOs de roubos consumados e total de ocorrências de roubo de celular na cidade.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Outras variáveis que apresentam alto coeficiente de correlação, que valem descrever: “DESCONHECIDO” — casos em que o bandido é desconhecido da vítima; “MOTO”, para os casos em que o bandido efetua o roubo pilotando uma moto; “PÚBLICA”, para os casos em que o roubo do celular ocorreu em uma via pública; “NOITE”, para os casos de roubo de celular ocorrido entre às 18H e 12H e “SAMSUNG”, marca de celular cujos registros apontam ser a mais roubada.

Dentre as variáveis acima citadas, todas elas apresentam uma relação linear entre o TOTAL de ocorrências de roubo de celular registrados na cidade de São Paulo.



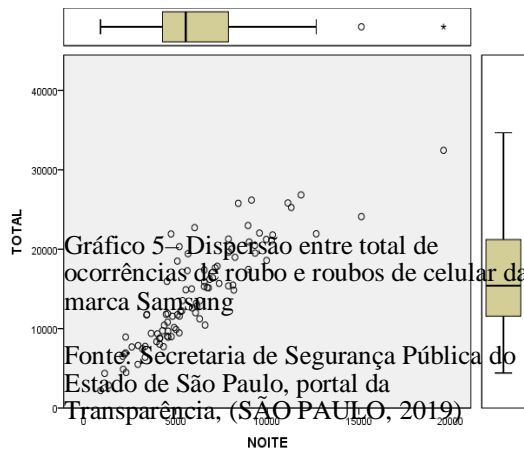


Gráfico 7– Dispersão entre total de ocorrências de roubo e roubos de celular à noite

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

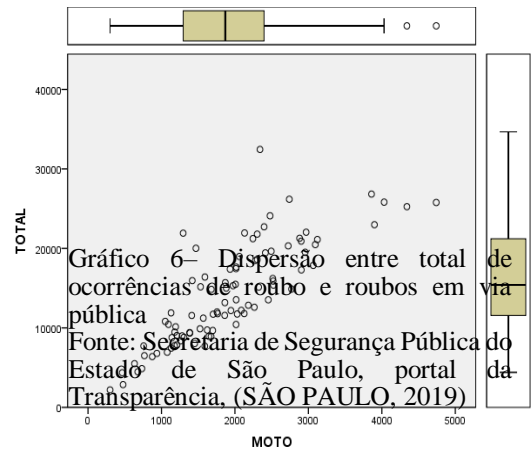


Gráfico 8– Dispersão entre total de ocorrências de roubo e roubos de celular em que o bandido estava em uma moto

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

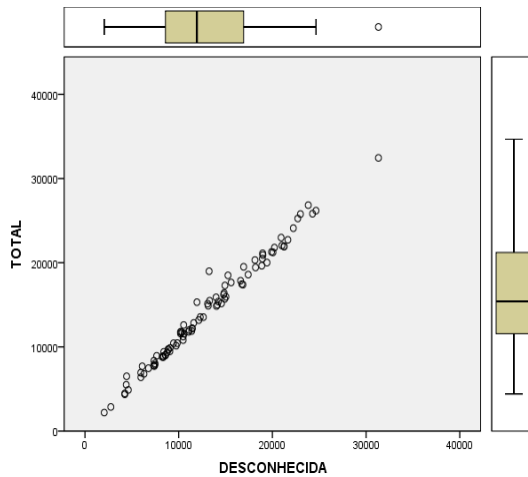


Gráfico 9– Dispersão entre total de ocorrências de roubo e roubos de celular em que o bandido não conhecia a vítima
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

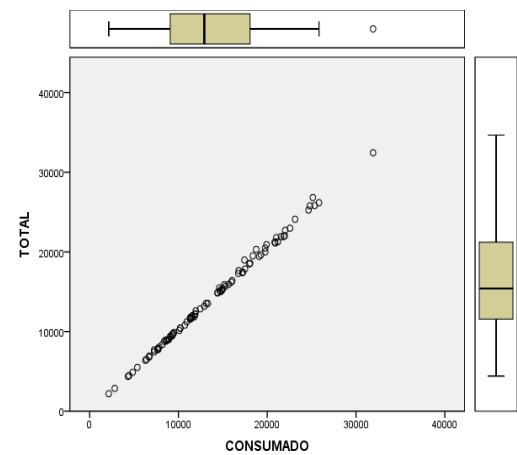


Gráfico 10– Dispersão entre total de ocorrências de roubo e roubos de celular consumados
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Ainda que cada uma das possíveis variáveis X apresentem forte Correlação de Pearson entre elas e a variável “TOTAL”, não é desejável que essas variáveis X tenham forte correlação entre si. Ou seja, para montar um modelo a fim de inferir o TOTAL de ocorrências de roubo de celular é necessário que essas variáveis X não estejam correlacionadas, o que seja que apresentem índice multicolinearidade abaixo de 30.

Mas antes desses testes, que serão feitos na próxima seção, pode utilizar-se a matriz de Correlação de Pearson abaixo. Os quadrados de bolas vermelhas formados nos cantos superiores e inferiores, dos lados direito e esquerdo, demonstram variáveis X correlacionadas.

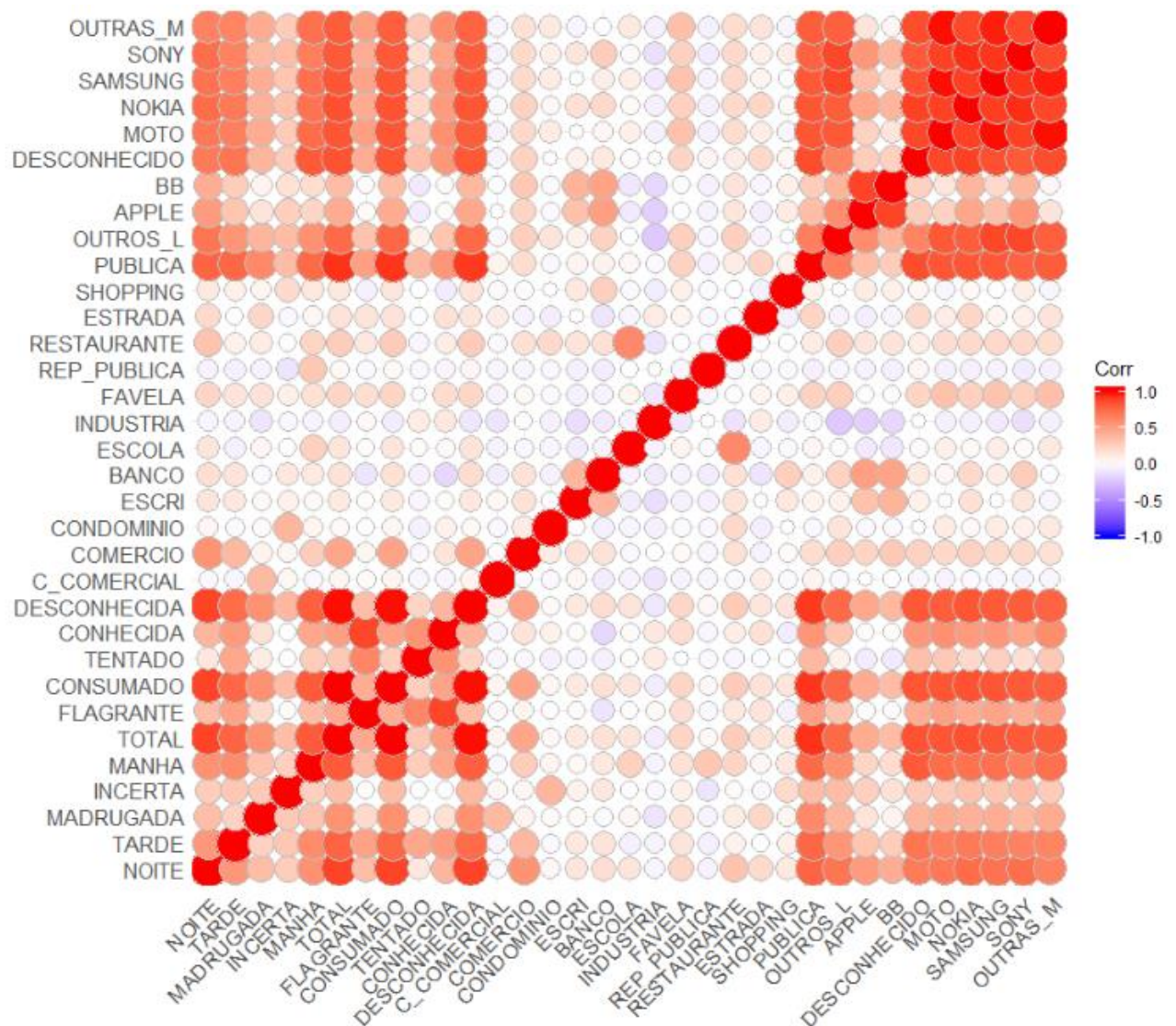


Gráfico 11– Matriz de correlação de roubos de celular na cidade de São Paulo.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.5 Regressão Linear

Em regressão linear múltipla com a técnica Stepwise do tipo forward foram testadas variáveis que poderiam ser capazes de prever o número total de ocorrências dos roubos de celular. De modo que o Y previsto é a variável “TOTAL”.

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$$

Diagrama explicativo da equação:

- \hat{Y}_i : Valor estimado (ou previsto) de Y
- b_0 : Intercepto estimado
- b_1, b_2, \dots, b_k : Coeficientes angulares estimados
- $X_{1i}, X_{2i}, \dots, X_{ki}$: Variáveis independentes

Sendo que as variáveis X a serem testadas demonstraram, a priori, atender alguns dos pré-requisitos necessários para uma Regressão confiável durante análises de estatística descritivas anteriormente realizadas:

a) estarem altamente correlacionadas ao Y, representado pelo total de ocorrências de roubo de celular;

b) apresentarem relação linear com o Y.

São elas, as variáveis “NOITE”, que informa o número de ocorrências em que o roubo do celular ocorreu entre às 18h e às 0h); “MOTO”, que indica o número de celulares roubados em que o bandido utilizava uma moto como veículo; “FLAGRANTE”, que trata-se da ocorrência que houve flagrante delito) e “SAMSUNG” (ocasiões em que a marca do celular roubado é da marca Samsung, a marca que obteve o maior número de registros);

Tabela 5– variáveis inseridas e removidas do modelo de regressão linear.

Variáveis Inseridas/Removidas^a

Modelo	Variáveis inseridas	Variáveis removidas	Método
1	NOITE ^b		Inserir
2	MOTO ^b		Inserir
3	FLAGRANTE ^b		Inserir
4	SAMSUNG ^b		Inserir

a. Variável Dependente: TOTAL

b. Todas as variáveis solicitadas inseridas.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Foram ajustados 4 modelos lineares, pelo tipo forward em que foram sendo adicionadas variáveis na ordem da tabela acima, seguindo a lógica decrescente do coeficiente de Correlação de Pearson. O R ao quadrado de todos os ajustes é alto, sendo o último modelo, que contém as 4 variáveis Y o maior deles, supostamente explicando 86,6% da variância do total de ocorrências de roubo de celular.

No entanto, comparados o R quadrado Ajustado, dos modelos o número 3 (com “NOITE”; “MOTO” e “FLAGRANTE”) e o número 2 (com “NOITE” e “MOTO”) aparecem empatados em primeiro lugar, ambos com 86,1%.

Do modelo 2 para o modelo 3 ganhou-se 1% de R ao quadrado. Porém, o erro padrão do modelo 3 é ligeiramente maior do que o do modelo 2. Essa mudança também não apresentou significância estatística, o valor p de mudança do modelo 3 é de 0,33.

O mesmo acontece com mudança do modelo 3 para o modelo 4, não há significância estatística com valor de 0,84. Portanto, os modelos 3 e 4 não são confiáveis para prever o total de ocorrências.

O coeficiente do teste de Durbin Watson de 1,8 indica que os resíduos não estão correlacionados entre si, ou seja, os resíduos são independentes, outro pré-requisito exigido para que o modelo de regressão seja confiável para prever o TOTAL de ocorrências.

Tabela 6– Comparação entre Modelos de Regressão Linear.

Resumo do Modelo ^e										
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Mudança de R quadrado	Estatísticas de mudança			Sig. Mudança F	Durbin-Watson
						Mudança F	df1	df2		
1	,877 ^a	,770	,767	3020,734	,770	303,843	1	91	,000	
2	,930 ^b	,864	,861	2330,204	,095	62,925	1	90	,000	
3	,930 ^c	,866	,861	2331,083	,001	,932	1	89	,337	
4	,930 ^d	,866	,860	2344,051	,000	,018	1	88	,894	1,820

a. P

b. Preditores: (Constante), NOITE, MOTO

c. Preditores: (Constante), NOITE, MOTO, FLAGRANTE

d. Preditores: (Constante), NOITE, MOTO, FLAGRANTE, SAMSUNG

e. Variável Dependente: TOTAL

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Já a análise de variância mostra que todos os modelos ajustados apresentam significância estatística, ou desempenho melhor do que a média aritmética para explicar o Y.

Tabela 7– Análise de Variância (ANOVA)

ANOVA ^a						
Modelo		Soma dos Quadrados	Df	Quadrado Médio	Z	Sig.
1	Regressão	2772518223,767	1	2772518223,767	303,843	,000 ^b
	Resíduo	830359991,416	91	9124835,071		
	Total	3602878215,183	92			
2	Regressão	3114191679,875	2	1557095839,938	286,766	,000 ^c
	Resíduo	488686535,307	90	5429850,392		

	Total	3602878215,183	92			
3	Regressão	3119257038,560	3	1039752346,187	191,344	,000 ^d
	Resíduo	483621176,623	89	5433945,805		
	Total	3602878215,183	92			
4	Regressão	3119355583,704	4	779838895,926	141,929	,000 ^e
	Resíduo	483522631,479	88	5494575,358		
	Total	3602878215,183	92			

a. Variável Dependente: TOTAL

b. Preditores: (Constante), NOITE

c. Preditores: (Constante), NOITE, MOTO

d. Preditores: (Constante), NOITE, MOTO, FLAGRANTE

e. Preditores: (Constante), NOITE, MOTO, FLAGRANTE, SAMSUNG

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Porém, dentre os coeficientes, os modelos 3 e 4 com as variáveis “FLAGRANTE” e “SAMSUNG”, indicam não ter significância estatística para compor o modelo, sendo, respectivamente 0,34 e 0,89. Ou seja, a hipótese de que a influência dessas variáveis é nula está bem acima de 5%.

Portanto, o modelo 2 é a melhor opção para explicar a variância do TOTAL de ocorrências de roubo de celular. De modo que, a variável “NOITE” tem 1,4 vezes a importância da variável “MOTO” para explicar ou prever a variação de TOTAL, conforme indicam os coeficientes padronizados.

Ainda que o Intercepto do modelo 2 -- que as duas variáveis restantes, “NOITE” e “MOTO” -- esteja acima de 5%, pois isso não é um problema por indicar apenas que a reta da contante (B0) pode passar por sua origem.

Como o valor de tolerância do modelo 2 é maior que 0,1 e de VIF é menor do que 10, pode-se afirmar que não existe multicolinearidade entre as variáveis X.

Tabela 8– Coeficientes dos 4 modelos de Regressão Linear múltipla ajustados

Coeficientes ^a												
Modelo	Coeficientes não padronizados		Coeficientes padronizados	t	Sig.	95,0% Intervalo de Confiança para B		Correlações			Estatísticas de colinearidade	
	B	Erro Erro				Limite inferior	Limite superior	Ordem zero	Parcial	Parcial e	Tolerância	VIF
1 (Constante)	3284,439	701,546		4,682	,000	1890,904	4677,973					
NOITE	1,789	,103	,877	17,431	,000	1,585	1,993	,877	,877	,877	1,000	1,000
2 (Constante)	956,971	615,596		1,555	,124	- 266,017	2179,960					
NOITE	1,205	,108	,591	11,140	,000	,990	1,419	,877	,761	,432	,536	1,866
MOTO	3,064	,386	,421	7,933	,000	2,297	3,832	,823	,641	,308	,536	1,866
3 (Constante)	972,488	616,037		1,579	,118	- 251,565	2196,542					
NOITE	1,204	,108	,590	11,129	,000	,989	1,419	,877	,763	,432	,536	1,866
MOTO	2,912	,418	,400	6,973	,000	2,082	3,741	,823	,594	,271	,459	2,178
FLAGRANTE	,422	,437	,043	,965	,337	-,447	1,291	,440	,102	,037	,758	1,320
4 (Constante)	984,625	626,059		1,573	,119	- 259,535	2228,785					
NOITE	1,207	,111	,592	10,859	,000	,986	1,428	,877	,757	,424	,513	1,948
MOTO	3,131	1,692	,430	1,851	,068	-,231	6,493	,823	,194	,072	,028	35,356
FLAGRANTE	,418	,441	,043	,947	,346	-,459	1,295	,440	,100	,037	,753	1,328
SAMSUNG	-,143	1,067	-,031	-,134	,894	-2,262	1,977	,825	-,014	-,005	,028	35,983

a. Variável Dependente: TOTAL

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2019)

Em termos de normalidade de resíduos, o Erro de Valor Previsto padronizado atende ao pré-requisito de estar entre -3 e 3, sinalizando não haver outliers. Enquanto o Erro Resíduo padronizado tem sua máxima de 4,7, indicando haver um Upper Outlier, questionando, portanto, a confiabilidade preditiva regressão.

A curva de distribuição dos resíduos não se trata de uma perfeita distribuição normal, mas seu formato torna possível considerá-la parcialmente, conforme indica o histograma.

Já o gráfico de resíduos padronizados indica, abaixo à direita, mostra esse Upper Outlier, circulado em vermelho e que essa regressão pode ter problemas de heterocedasticidade, ou seja,

os resíduos podem não apresentar variância constante. É possível perceber através do gráfico não apresenta distribuição aleatória.

Tabela 9– Análise de resíduos.

Estatísticas de resíduos^a

	Mínimo	Máximo	Média	Erro Desvio	N
Valor previsto	2963,28	31604,89	14226,53	5822,889	93
Resíduo	-2864,840	11232,712	,000	2292,527	93
Erro Valor previsto	-1,934	2,984	,000	1,000	93
Erro Resíduo	-1,222	4,792	,000	,978	93

a. Variável Dependente: TOTAL

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2019)

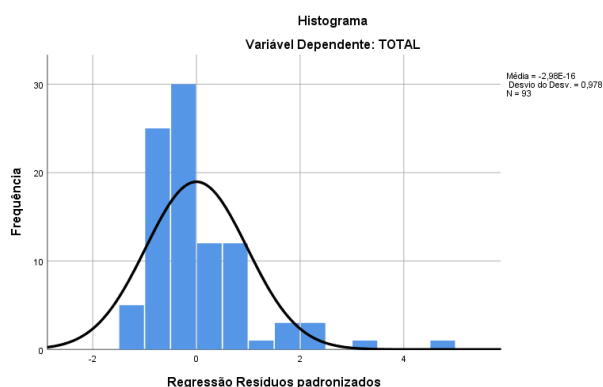


Gráfico 12– Normalidade de resíduos.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

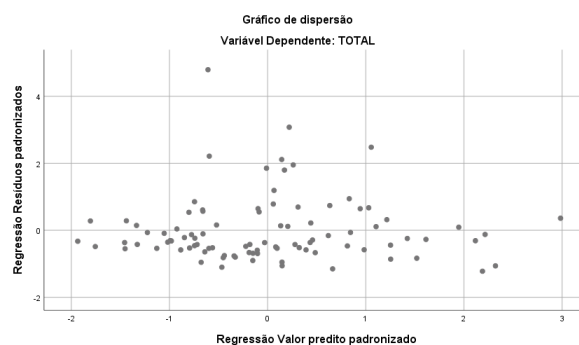


Gráfico 13– Resíduos padronizados Vs valor previsto padronizado.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Como meio de inferência acerca do “TOTAL” (de ocorrências de roubo de celular) o modelo que apresenta melhor desempenho é o modelo 2-- que inclui o número de roubos de celular em que o bandido estava em uma “MOTO” e no período “NOITE” (número de roubos de aparelhos celulares entre às 18h e às 00h). Sendo modelo estatisticamente significativo:

$$[F(2,90)= 286,766;p <0,001; R^2 = 0,864].$$

As ocorrências de roubo de celular à NOITE ($\beta=0,591$, $t=11,140$; $p <0,001$)

e as ocorrências sobre uma MOTO ($\beta= 0,421$, $t=7,933$; $p <0,001$) pode ser preditores do TOTAL de boletins de ocorrência registrados. A equação que descreve essa relação é

$$y=b_0 +b_1 \cdot x_1+ b_2 \cdot x_2$$

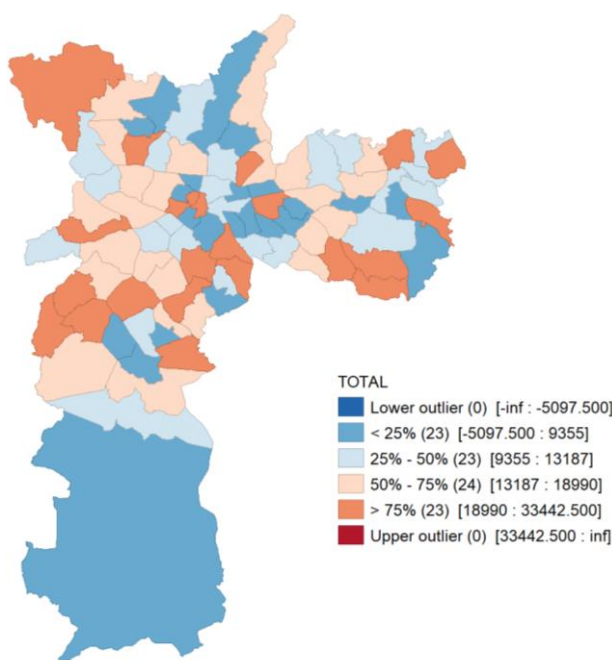
$$\text{“TOTAL”} = 956,971 \cdot (\text{“NOITE”}) + 3,064 \cdot (\text{“MOTO”}).$$

PARTE II- MODELOS ESPACIAIS

3.6 Geral e descritivas

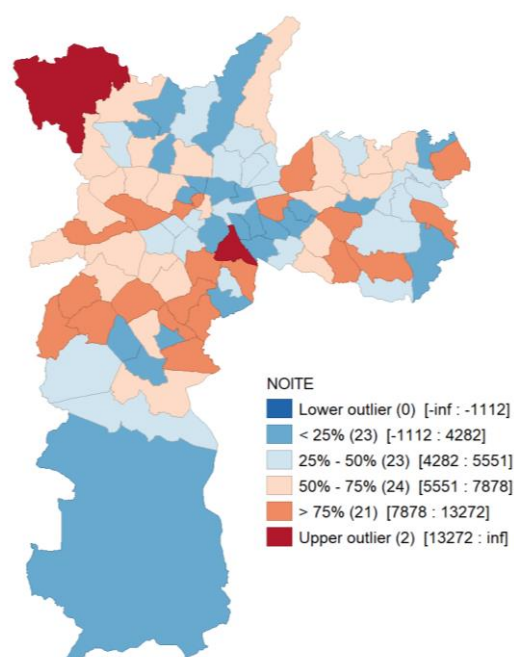
O Box Map (Boxplot em formato de mapa) indicam que não há outliers, nem para baixo nem para cima, em relação ao “TOTAL” de ocorrências de roubos de celular registradas na cidade(Y). Já as ocorrências de roubo de celular no período da “NOITE”, uma das variáveis X incluídas nos modelos, indicam que os Distritos Policiais do Ipiranga e de Perus têm o maior número de roubos no período noturno a ponto de estarem fora de padrão para cima (Upper Outlier).

As ocorrências de roubo em que o meliante estava se locomovendo com uma “MOTO”, outra variável X dos modelos, tem como Upper Outliers os registros do DP de São Mateus e de Itaim Paulista. Os registros de roubo da marca “SAMSUNG”, marca com maior número de registros, têm cinco Distritos Policias Jabaquara, São Mateus, Capão Redondo e Itaim Paulista.



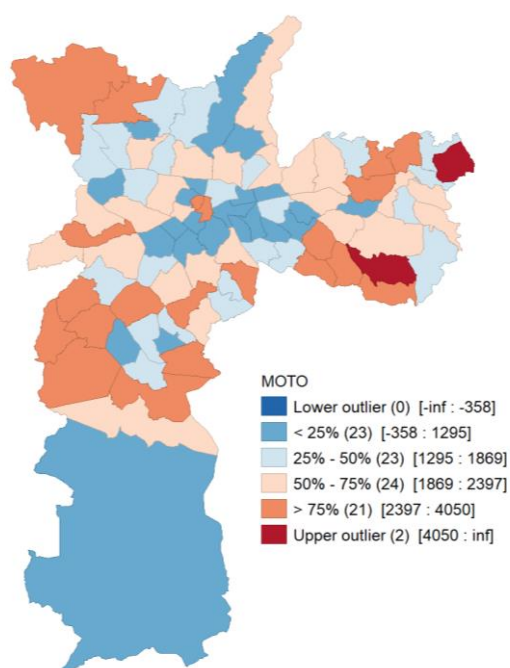
Mapa 3– Boxmap Totalidade de ocorrências de roubo de celular na cidade de São Paulo, desde 2010 a 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



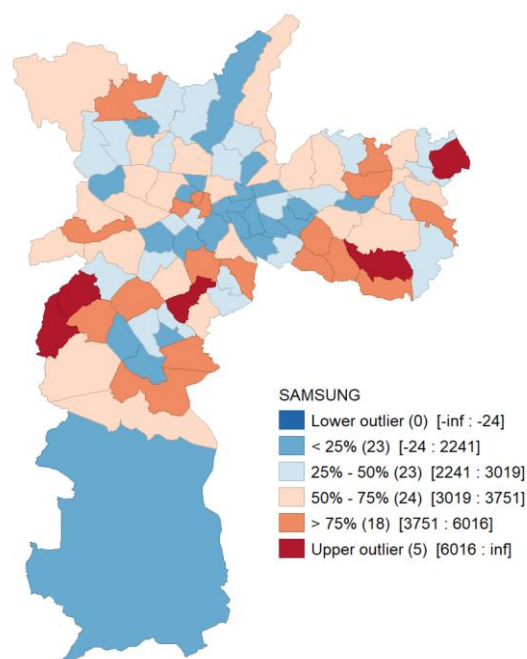
Mapa 4– Boxmap Totalidade de ocorrências de roubo de celular à noite na cidade de São Paulo, desde 2010 a 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 5– Boxmap Totalidade de ocorrências de roubo de celular na cidade de São Paulo, desde 2010 a 2018, em que o bandido estava em uma moto.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

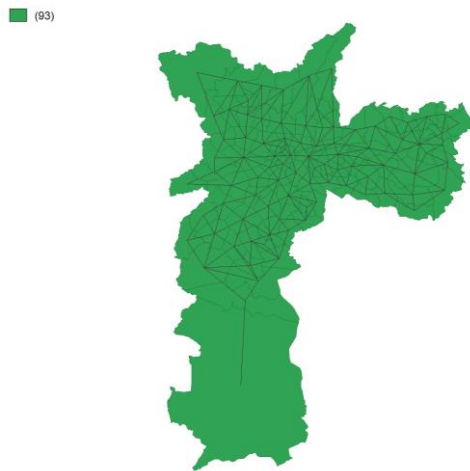


Mapa 6– Boxmap Totalidade de ocorrências de roubo de celular da marca Samsung na cidade de São Paulo, desde 2010 a 2018.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.6.1 Matriz de vizinhança

Foram testados 2 modelos espaciais com as mesmas variáveis, a partir de uma matriz de vizinhança do tipo Queen, de Ordem 1, considerando apenas o primeiro vizinho. Conforme se observa no histograma abaixo, a distribuição de vizinha entre os 93 Distritos Policiais se assemelha à curva normal.



Mapa 7– matriz de vizinhança.

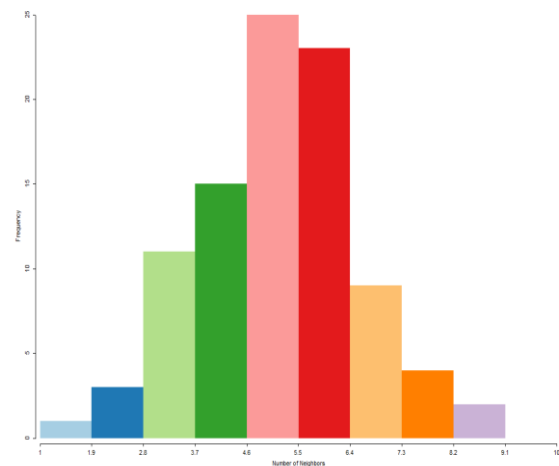


Gráfico 14– histograma de vizinhança

3.6.2 Autocorrelação espacial

Índice de Global Moran (I) representa autocorrelação considerando apenas o vizinho, sendo uma média atributo “TOTAL” (de ocorrências de roubo de celular na cidade de São Paulo) na cidade de São Paulo em estudo e W_{ij} os elementos da matriz normalizada, dada pela equação:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^n (Z_i - \bar{Z})^2}$$

O I de Moran indica o quanto daquela função é explicada pelo espaço, ou seja, a influência dos vizinhos na respectiva variável em análise. No caso de “TOTAL” apenas 5,8% de sua variância pode ser explicada pelo espaço geográfico; de “NOITE” 8,9% de “MOTO” 18,9%.

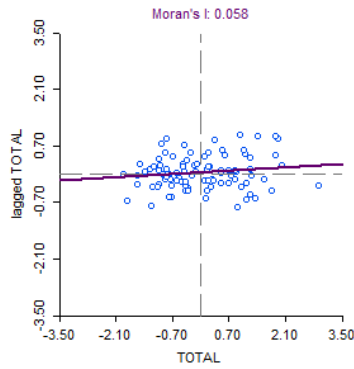


Gráfico 15– Moran Scatter Plot do total de ocorrências de roubo de celular.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

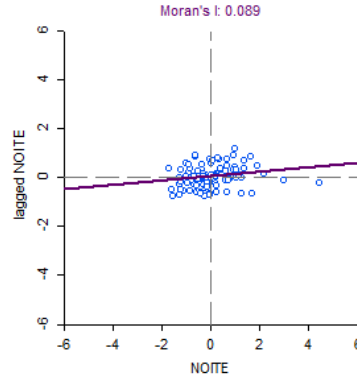


Gráfico 16– Moran Scatter Plot das ocorrências de roubo de celular à noite

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

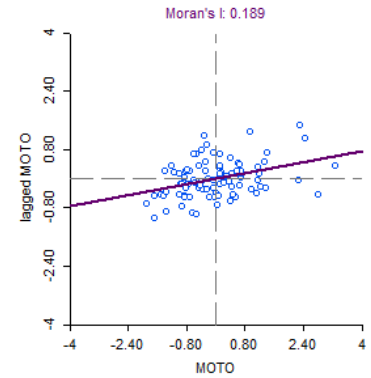


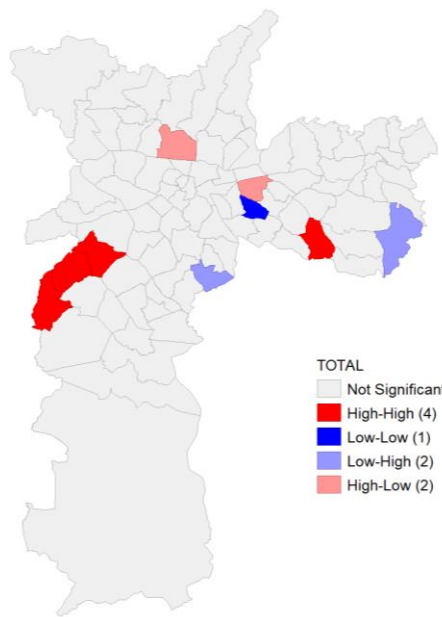
Gráfico 17– Moran Scatter Plot das ocorrências de roubo de celular em que o bandido estava em uma moto.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

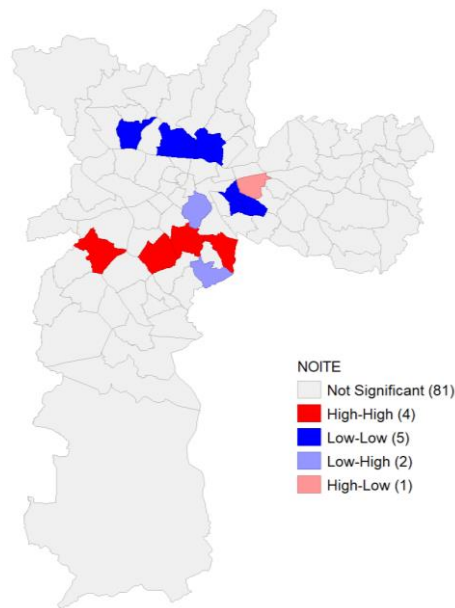
Já o Índice Local de Moran (I_i) representa o valor da correlação espacial específico para cada área i , a partir dos valores normalizados Z_i do atributo “TOTAL”, sendo dado por:

$$I_i = \frac{Z_i \sum_{j=1}^n W_{ij} Z_j}{\sum_{j=1}^n Z_j^2}$$

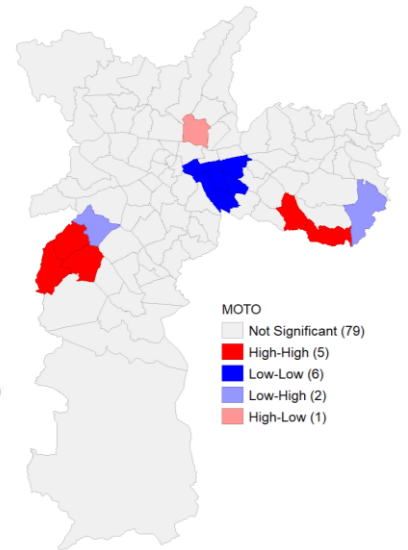
Os pontos em vermelho Lisa Cluster Map (High-High) indicam os distritos em que o Distrito Policial e seus vizinhos apresentam alto número de Ocorrências para cada uma das variáveis abaixo. Os pontos em Azul (Low-Low) indicam os Distrito Policiais em que há o menor número de ocorrências e seus vizinhas. Já os mapas de significância indicam os Distritos Policiais em que há significância estatística para análise espacial. Que para as 3 variáveis descritas variam de 9 a 14 DP apenas.



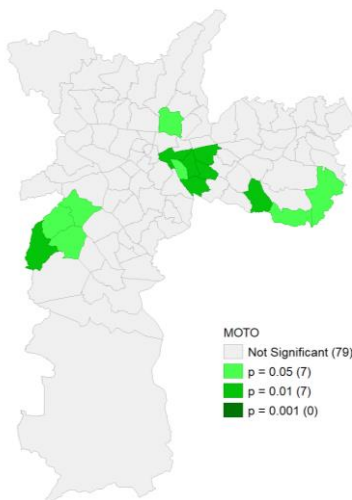
Mapa 8– Lisa Cluster Map total de ocorrências de roubo de celular. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



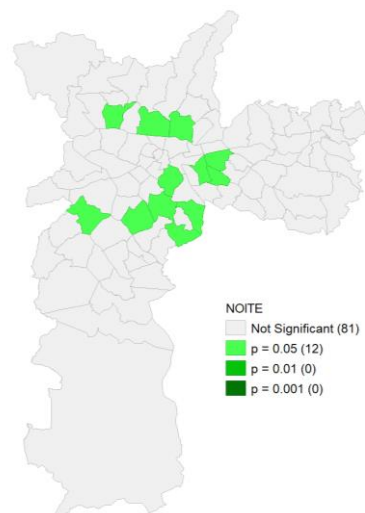
Mapa 9– Lisa Cluster Map das ocorrências de roubo de celular à noite. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



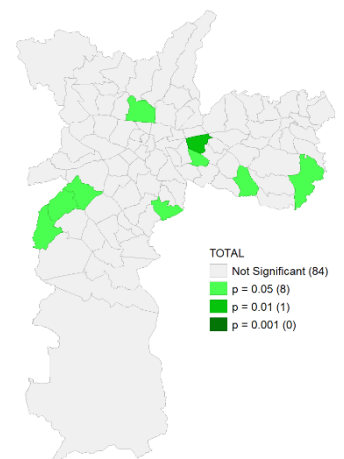
Mapa 10– Lisa Cluster Map das ocorrências de roubo de celular na cidade de São Paulo, desde 2010 a 2018, em que o bandido estava em uma moto. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 11– Significance Map das ocorrências de roubo de celular na cidade de São Paulo, desde 2010 a 2018, em que o bandido estava em uma moto. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 12– Significance Map das ocorrências de roubo de celular à noite. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 13– Significance Map do total de ocorrências de roubo de celular. Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.6.3 Regressão Espacial Global

Em comparação ao modelo linear, anteriormente, descrito os modelos espaciais não apresentam grandes ganhos de inferência. O R^2 , o quanto da variância de Y, é explicada pelas variáveis X. Há ganho, porém ele baixo não chega a 1%.

No caso do modelo Espacial Linear há uma perda no Critério de formação de Akaike, que se torna maior do que o modelo de Regressão Linear. Já o modelo Spatial Error, que analisa o padrão dos resíduos apresenta ganhos tanto em relação ao R^2 , quanto em diminuição de Akaike.

Tabela 10– Comparação entre as Regressões espaciais e convencional.

Tipo de Modelo	Akaike info criterion	R^2
Linear	1709.06	0.861
Spatial Lag	1709.98	0.866
Spatial Error	1707.61	0.867

Outro ponto de vantagem na utilização do modelo espacial em relação à Regressão Linear é que tanto no Spatial Lag, quanto no Spatial Error, a equação não apresenta problemas de heterocedasticidade. Portanto, a variância dos resíduos é constante, uma vez que o teste de Breusch-Pagan é de 0,822 para Spatial Lag e de 0,789 para Spatial Error, refutando-se a hipótese nula de que há variância inconstante de resíduos.

O modelo Spatial Error é o mais vantajoso dentre os espaciais testados e sua equação inclui um valor de lambda se reflete uma penalidade à complexidade do modelo, que, porém, não refuta hipótese nula de que ela passe pelo 0.

Tabela 11– Resultados Spatial error.

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	784.447	636.521	1.2324	0.21780
NOITE	1.23415	0.10727	11.5051	0.00000
MOTO	3.04817	0.390922	7.79738	0.00000
LAMBDA	0.20348	0.145794	1.39567	0.16281

REGRESSION DIAGNOSTICS

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	2	0.4717	0.78988

DIAGNOSTICS FOR SPATIAL DEPENDENCE

SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : Distritos_Policiais_SHP1

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	1.4530	0.22805

COEFFICIENTS VARIANCE MATRIX

CONSTANT	NOITE	MOTO	LAMBDA
405159.552845	-14.690444	-118.883033	0.000000
-14.690444	0.011507	-0.028775	0.000000
-118.883033	-0.028775	0.152820	0.000000
0.000000	0.000000	0.000000	0.021256

Outra vantagem do Modelo Spatial Error sobre o Modelo de Regressão Linear é que diminui a Auto-correlação Espacial Global (Índice de Moran) entre os resíduos. Passando de 7,5% para 1,2%.

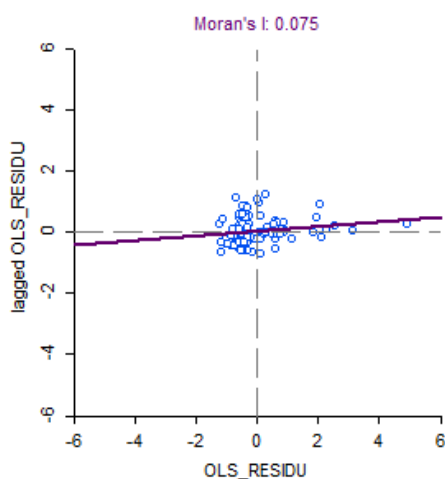


Gráfico 18– Moran Scatter Plot dos resíduos da regressão convencional
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

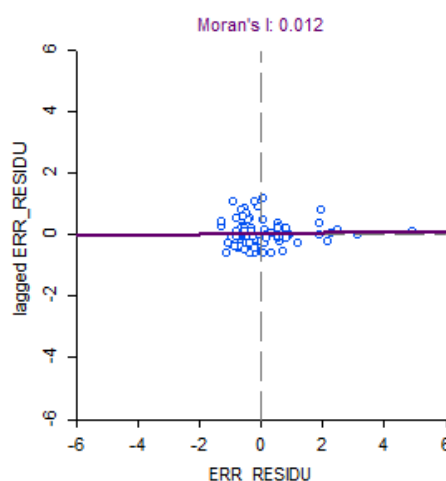
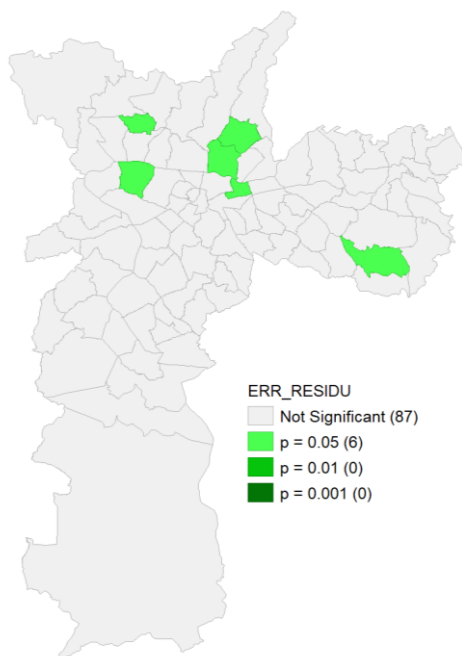


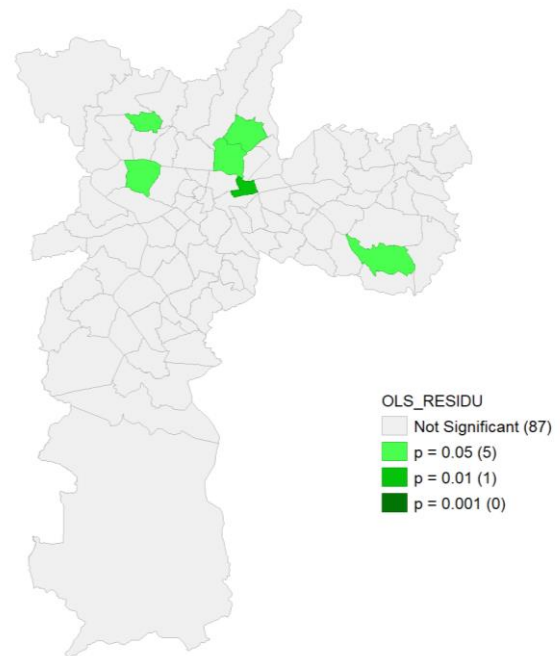
Gráfico 19– Moran Scatter Plot dos resíduos da regressão espacial (Spatial Error)
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

O modelo Spatial Error também diminuiu à auto-correlação espacial local (I Moran Local) dos resíduos. No modelo comum, apenas 2 dos 93 Distritos Policiais apresentam resíduos com alto I de Moran Local (DP do Belém e DP de São Mateus), enquanto que no Modelo Espacial (Spatial Error) apenas o distrito do Belém se manteve com alto I de Moran Local, descrito em vermelho.

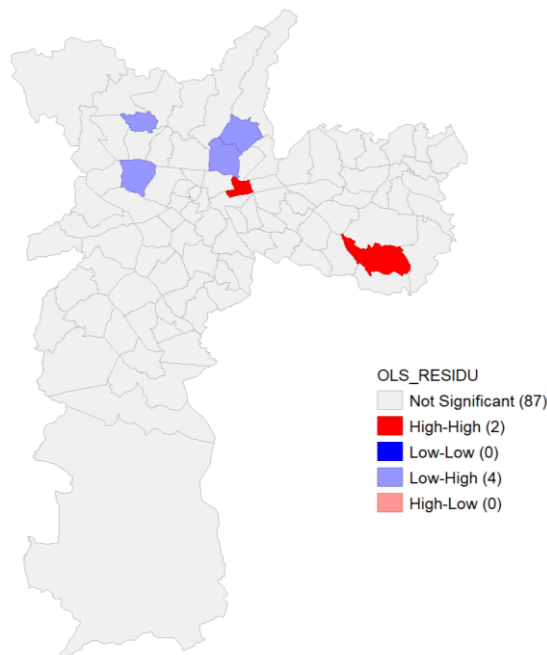
Em termos de significância os resíduos do modelo linear e do Spatial Error Model permaneceram parecidos, apenas 6 Distritos Policiais apresentaram alto I de Moran dos resíduos.



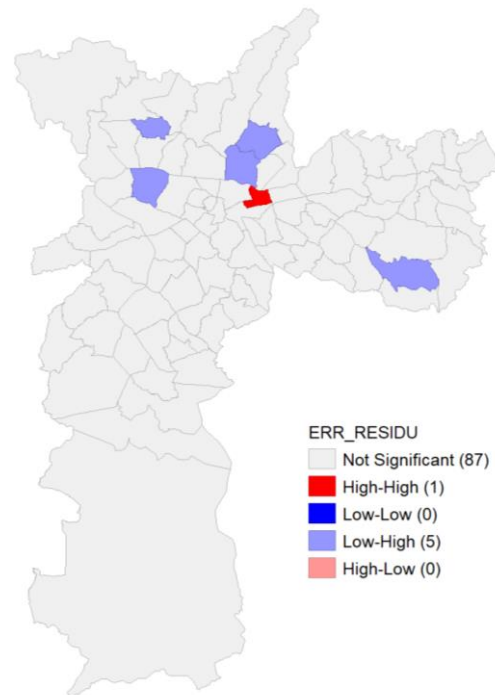
Mapa 14– Significance Map dos resíduos da regressão espacial (Spatial Error)
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 15– Significance Map dos resíduos da regressão convencional
Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 16– Lisa Cluster Map dos resíduos da regressão convencional
 Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)



Mapa 17– Lisa Cluster Map dos resíduos da regressão espacial (Spatial Error)
 Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

Ainda que haja ganhos de acurácia, observando o Multiplicador de Lagrange nenhum dos modelos espaciais apresenta valores de significância estatística com valor p menor que 5%. O mais baixo deles é o Lagrange Robusto do Spatial Error.

Ademais, a auto correlação dos resíduos do modelo convencional é de apenas 7,4%, conforme apontado anteriormente. Na prática, isso significa que o modelo linear convencional, descrito no capítulo anterior.

Tabela 12– Diagnósticos de análise espacial- Multiplicadores de Lagrange.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
 FOR WEIGHT MATRIX : Distritos_Policiais_SHP1
 (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.0747	1.3694	0.17087
Lagrange Multiplier (lag)	1	0.8815	0.34779
Robust LM (lag)	1	3.0376	0.08136
Lagrange Multiplier (error)	1	1.2236	0.26865
Robust LM (error)	1	3.3797	0.06600
Lagrange Multiplier (SARMA)	2	4.2612	0.11877

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

PARTE III- MACHINE LEARNING

4 CHANCES DE FLAGRANTE

4.1 Geral

Nessa segunda parte do presente artigo, foram utilizadas técnicas de aprendizado de máquina e treino de algoritmos para analisar cada uma das mais de 1 milhão de ocorrências da variável “FLAGRANTE”, que foram transformadas de “sim” e “não” para variáveis Dummy.

Portanto, cada uma das ocorrências em que houve Flagrante Delito de roubos de celular o “Sim” tornou-se “1”, do mesmo modo que observações dos Boletins de Ocorrência em que havia descrição “Não” foram transformadas em “0”.

O objetivo é treinar para encontrar os algoritmos para encontrar o que apresenta a melhor previsibilidade dos casos de flagrante e observar também o que apresentam menor erro total. De acordo com o Art. 302, do Código Penal brasileiro, considera-se flagrante, quanto:

“I - está cometendo a infração penal; II - acaba de cometê-la; III - é perseguido, logo após, pela autoridade, pelo ofendido ou por qualquer pessoa, em situação que faça presumir ser autor da infração; IV - é encontrado, logo depois, com instrumentos, armas, objetos ou papéis que façam presumir ser ele autor da infração. Art. 303. Nas infrações permanentes, entende-se o agente em flagrante delito enquanto não cessar a permanência.” (ART.302 e 303, CÓDIGO PENAL, 1940).

Além do Flagrante foram analisadas outras 6 variáveis, que descrevem o Local que ocorreu o roubo do celular; o período; a marca do celular; se o autor do crime era conhecido da vítima e se o roubo foi consumado.

Tabela 13– variáveis utilizadas para análise de flagrante

[1]	"PERIDOOCCORRENCIA"	"BO_AUTORIA"	"FLAGRANTE_N"
[4]	"DESCRICAOLocal"	"RUBRICA"	"STATUS"
[7]	"MARCA_CELULAR"		

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

4.1.1 Treino e Teste

Foram selecionados aleatoriamente 80% dos dados para treino, usados para o ajuste dos modelos. Enquanto os demais 20% (que o modelo não tem acesso durante o treino) como Teste, para aplicação da fórmula ou do modelo a fim de testar a acurácia do mesmo.

Tabela 14– equação para estabelecer conjuntos de treino e teste.

```
set.seed(1)
treino_f <- sample(nrow(flagrante), size = 0.8*nrow(flagrante))
```

4.2 Regressão Logística

Para a Regressão logística, que é do tipo binomial, (dada pela fórmula abaixo) foram treinados algoritmos em todas as observações.

$$y = \frac{e^{ax+b}}{1 + e^{ax+b}}$$

Tabela 15– fórmula da regressão logística e estatísticas descritivas do Y.

```
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Call:
```

```
glm(formula = FLAGRANTE_N ~ ., family = "binomial", data = flagrante[treino_f,
])
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-2.252	0.000	0.000	0.000	2.521

Os coeficientes abaixo com dois ou com três asteriscos ao lado apresentam significância estatística para o modelo. Por exemplo, as ocorrências em Shoppings centers, em estações de metro apresentam a probabilidade de 99% (valor p de 0, 001) de influenciarem na variância da ocorrência ou não de flagrante.

Do mesmo modo em que os roubos de celular cometidos em escritório, estabelecimento comercial, estabelecimento prisional não tem relação com a probabilidade de roubo de aparelho ser flagrado. O mesmo ocorre com os crimes em que o autor é desconhecido da vítima ou que o aparelho é da HTC ou é um Nextel.

Tabela 16– coeficientes da regressão logística

Coefficients:	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.065e+00	2.079e-01	5.122	3.02e-07 ***
PERIDOOCCORRENCIAA TARDE	3.243e-04	2.427e-02	0.013	0.989339
PERIDOOCCORRENCIADE MADRUGADA	2.689e-01	2.368e-02	11.352	< 2e-16 ***
PERIDOOCCORRENCIAEM HORA INCERTA	-3.558e+00	1.020e+00	-3.489	0.000484 ***
PERIDOOCCORRENCIAPELA MANHÃ	6.872e-02	2.279e-02	3.016	0.002563 **
BO_AUTORIADesconhecida	-2.636e+01	1.213e+02	-0.217	0.827969
BO_AUTORIAIgnorado	1.301e+00	4.822e-02	26.983	< 2e-16 ***
DESCRICAOLOCALCentro Comerc./Empresarial	1.924e-02	3.226e-01	0.060	0.952448
DESCRICAOLOCALComércio e serviços	-4.563e-01	2.036e-01	-2.241	0.025002 *
DESCRICAOLOCALCondominio Comercial	-2.119e+01	2.896e+03	-0.007	0.994161
DESCRICAOLOCALEntidade assistencial	-3.487e+00	1.046e+00	-3.334	0.000857 ***
DESCRICAOLOCALEscritório	-3.675e-01	2.643e-01	-1.390	0.164446
DESCRICAOLOCALEstabelecimento bancário	-1.929e-01	3.102e-01	-0.622	0.534022
DESCRICAOLOCALEstabelecimento de ensino	-2.681e+00	3.992e-01	-6.716	1.86e-11 ***
DESCRICAOLOCALEstabelecimento industrial	-1.801e+00	5.232e-01	-3.443	0.000576 ***
DESCRICAOLOCALEstabelecimento prisional	2.117e+00	5.596e+04	0.000	0.999970
DESCRICAOLOCALEstrada de ferro	-5.174e-01	1.465e+00	-0.353	0.724010
DESCRICAOLOCALFavela	-6.235e-01	3.799e-01	-1.641	0.100774
DESCRICAOLOCALHospedagem	-7.310e-01	3.448e-01	-2.120	0.033988 *
DESCRICAOLOCALLazer e recreação	-7.727e-01	2.495e-01	-3.097	0.001955 **
DESCRICAOLOCALLocal clandestino/ilegal	1.227e+01	1.549e+02	0.079	0.936878
DESCRICAOLOCALOutros	5.945e-02	2.122e+02	0.000	0.999776
DESCRICAOLOCALRepartição Pública	-7.137e-01	2.725e-01	-2.619	0.008818 **
DESCRICAOLOCALRestaurante e afins	-9.891e-01	2.139e-01	-4.625	3.74e-06 ***
DESCRICAOLOCALRodovia/Estrada	3.423e-01	2.496e-01	1.371	0.170302
DESCRICAOLOCALSaúde	-1.186e+00	2.725e-01	-4.350	1.36e-05 ***
DESCRICAOLOCALServiços e bens públicos	-2.494e+00	1.214e+00	-2.055	0.039875 *
DESCRICAOLOCALShopping Center	1.446e+00	3.869e-01	3.738	0.000185 ***
DESCRICAOLOCALSíndico	1.994e+00	2.289e+04	0.000	0.999930
DESCRICAOLOCALTemplo e afins	-1.970e+00	8.376e-01	-2.352	0.018694 *
DESCRICAOLOCALTerminal/Estação	-1.538e+00	2.438e-01	-6.311	2.78e-10 ***
DESCRICAOLOCALUnidade rural	-2.162e+01	7.443e+03	-0.003	0.997683
DESCRICAOLOCALVia pública	-7.154e-01	1.994e-01	-3.588	0.000333 ***
RUBRICA COLETIVO	3.256e-01	2.820e-01	1.155	0.248175
RUBRICA CONDOMINIO COMERCIAL	-2.023e+01	7.245e+03	-0.003	0.997773
RUBRICA CONDOMINIO RESIDENCIAL	1.181e+01	1.233e+02	0.096	0.923724
RUBRICA ESTABELECIMENTO	-3.458e-01	2.187e-01	-1.581	0.113930
RUBRICA ESTABELECIMENTO BANCARIO	-6.299e-01	3.962e-01	-1.590	0.111925
RUBRICA ESTABELECIMENTO COMERCIAL	-5.847e-02	7.813e-02	-0.748	0.454221
RUBRICA ESTABELECIMENTO ENSINO	-6.921e-03	8.263e-01	-0.008	0.993316
RUBRICA INTERIOR DE VEICULO	-7.066e-01	7.639e-02	-9.250	< 2e-16 ***
RUBRICA INTERIOR ESTABELECIMENTO	-3.511e-01	1.083e-01	-3.243	0.001182 **
RUBRICA INTERIOR TRANSPORTE COLETIVO	1.776e-01	1.381e-01	1.286	0.198374
RUBRICA JOALHERIA	-2.252e+01	2.839e+04	-0.001	0.999367
RUBRICA OUTROS	-6.261e-01	7.070e-02	-8.855	< 2e-16 ***
RUBRICA RESIDENCIA	3.866e-01	1.161e-01	3.330	0.000868 ***
RUBRICA TRANSEUNTE	-4.827e-01	5.925e-02	-8.146	3.76e-16 ***
RUBRICA VEICULO	-1.257e-01	5.693e-02	-2.208	0.027256 *
RUBRICADESCONHECIDA	-6.038e-02	5.589e-02	-1.080	0.280007

```

STATUSTentado      3.204e-01  6.905e-02  4.641 3.47e-06 ***
MARCA_CELULARBlackberry      -8.219e-02  9.248e-02 -0.889 0.374151
MARCA_CELULARDesconhecido      -5.235e-01  3.564e-02 -14.689 < 2e-16 ***
MARCA_CELULARHTC      -8.236e-01  3.259e-01 -2.527 0.011508 *
MARCA_CELULARHTC      3.066e+00  7.946e+04  0.000 0.999969
MARCA_CELULARHuawei      6.104e-01  1.746e-01  3.497 0.000471 ***
MARCA_CELULARLenovo      3.336e-01  1.195e-01  2.792 0.005239 **
MARCA_CELULARLG      1.172e+01  1.090e+02  0.108 0.914390
MARCA_CELULARMicrosoft      6.931e-01  1.858e-01  3.731 0.000191 ***
MARCA_CELULARMotorola      -5.165e-02  3.227e-02 -1.601 0.109465
MARCA_CELULARNextel      -8.991e-02  1.375e-01 -0.654 0.513080
MARCA_CELULARNokia      -1.197e-01  3.982e-02 -3.005 0.002652 **
MARCA_CELULAROutros      9.632e-02  3.431e-02  2.807 0.004999 **
MARCA_CELULARPositivo      -2.241e-01  1.545e-01 -1.451 0.146799
MARCA_CELULARSamsung      -2.947e-02  2.980e-02 -0.989 0.322588
MARCA_CELULARSiemens      -2.180e-01  4.936e-01 -0.442 0.658671
MARCA_CELULARSony      4.995e-02  5.460e-02  0.915 0.360300
MARCA_CELULARTOSHIBA      2.659e+00  2.281e+04  0.000 0.999907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

O gráfico a seguir representa o melhor ponto de corte para o modelo, que apresenta o menor o erro.

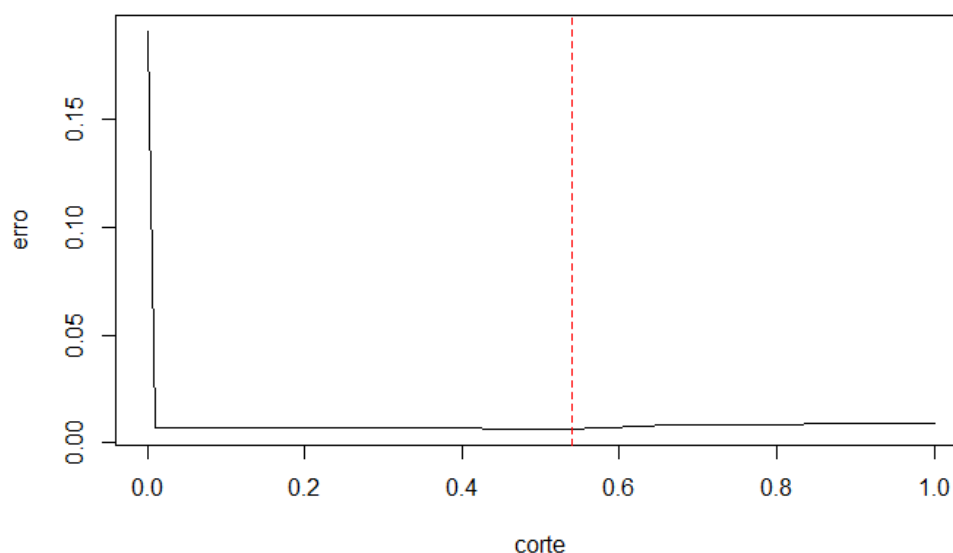


Gráfico 20— ponto de corte da regressão logística

A matriz de confusão do modelo logístico mostra que o algoritmo acertou 174.590 casos de não flagrante e errou apenas 1891. Assim como, acertou 6273 casos de flagrante e errou 4104.

Em termos de erro total o modelo erra em 5.995 (3,21%) e acerta 186.858 (96,79%). Para o gestor da política pública a informação mais relevante é o desempenho preditivo dos casos de flagrante, e, nesse quesito, após o treinamento, o algoritmo previu 60,4% das ocorrências de flagrante do conjunto de teste, conseguindo acertar 6273 casos dentre os 10.377.

Tabela 17– Matriz de confusão-Regressão Logística

predito		
observado	0	1
0	174590	4104
1	1891	6273

O ponto vermelho destaca o ponto ótimo da curva ROC, onde há menor erro total.

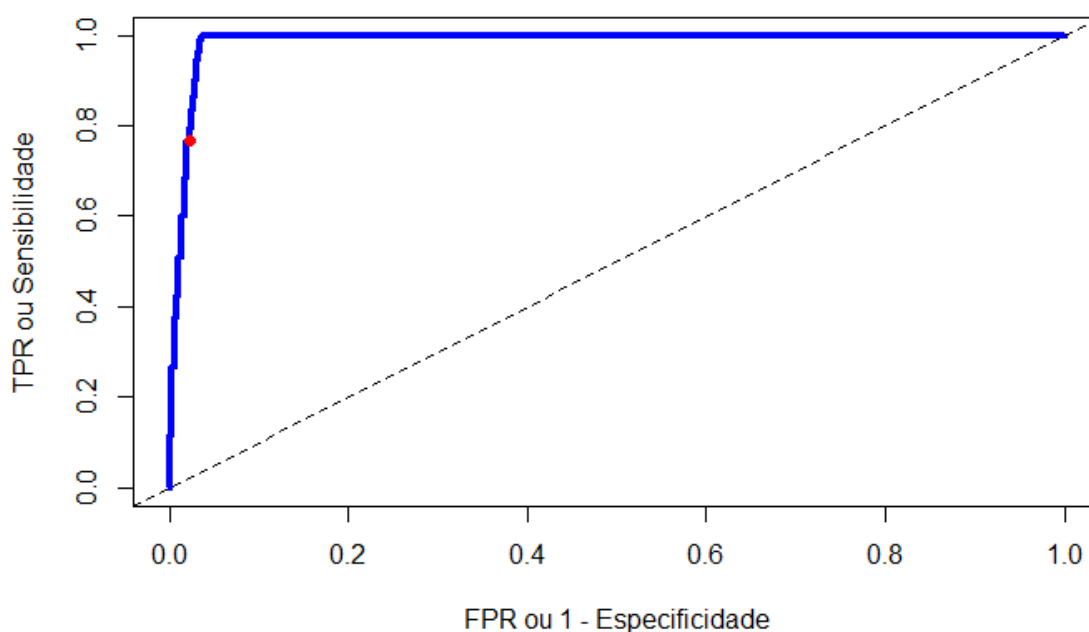


Gráfico 21– Curva ROC- Regressão Logística

4.3 Árvore de Decisão

As árvores de decisão e seus ensembles são algoritmos comumente utilizadas por alguns mercados para prever fraudes. As arvores também apresentam boa inferência, apesar de desempenho preditivo mais baixo que os demais algoritmos do mesmo tipo.

Essa árvore indica 92% de probabilidade de não haver um flagrante quando o autor do crime é desconhecido da vítima e quando o roubo ocorre num condomínio comercial.

A chance de registro do flagrante é de 8% se a pessoa for conhecida pela vítima. Se o roubo do celular for em uma joalheria ou em um estabelecimento comercial e o ladrão também for conhecido da vítima, a chance de registro do flagrante é de 6%.

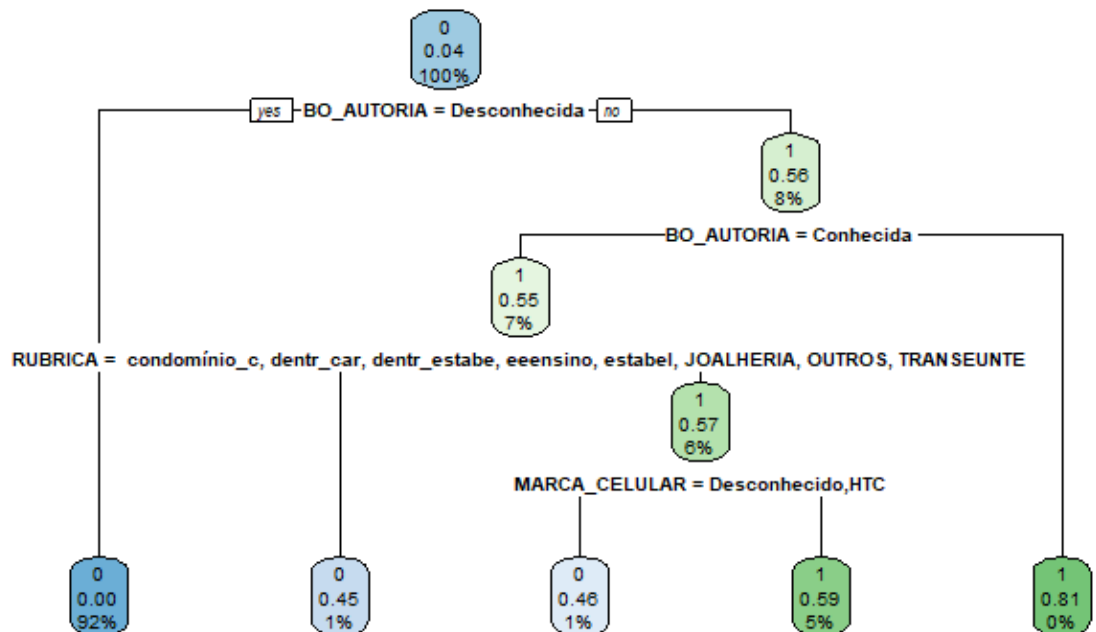


Gráfico 22– Árvore de decisão- Flagrante

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

A matriz de confusão apresenta um erro total maior do que a regressão logística: de 3,35%. O algoritmo disse e acertou 174.493 casos de não flagrante. e acertou 59,9% (6301) casos de flagrante.

Tabela 18– Matriz de confusão para árvore de decisão

pred_arvore_f		
0	1	
0	174493	4201
1	1863	6301

3.5 Bootstrap aggregating (Bagging)

Proposto por Leo Breiman, em 1994, o Bagging é uma técnica, cujo propósito é reduzir a variância de um método de aprendizagem e ajuda a reduzir o overfitting – ajuste excessivo que perde desempenho preditivo para dados novos ou para o conjunto de treino.

“is a technique for reducing the variance of an estimated prediction function. Bagging seems to work especially well for high-variance, low-bias procedures, such as trees. For regression, we simply fit the same regression tree many times to bootstrapped versions of the training data, and average the result. For classification, a committee of trees each cast a vote for the predicted class” (p.606, Elements of Statistical Learning, 2008)

Uma forma de reduzir a variância é por meio da utilização de muitos conjuntos—

Método de Ensembles—, treinando regressores em cada um dos conjuntos. Nesse caso, a resposta é resultante da média do conjunto, dado pela equação:

$$\hat{\psi}_{\text{bag}}(x) = \frac{1}{B} \sum_{i=1}^B \hat{\psi}_i(x).$$

Foram treinadas e testadas 500 árvores, sendo o melhor ajuste com menor erro para uso de 390 árvores, acima dessa quantidade a taxa de erro volta a crescer.

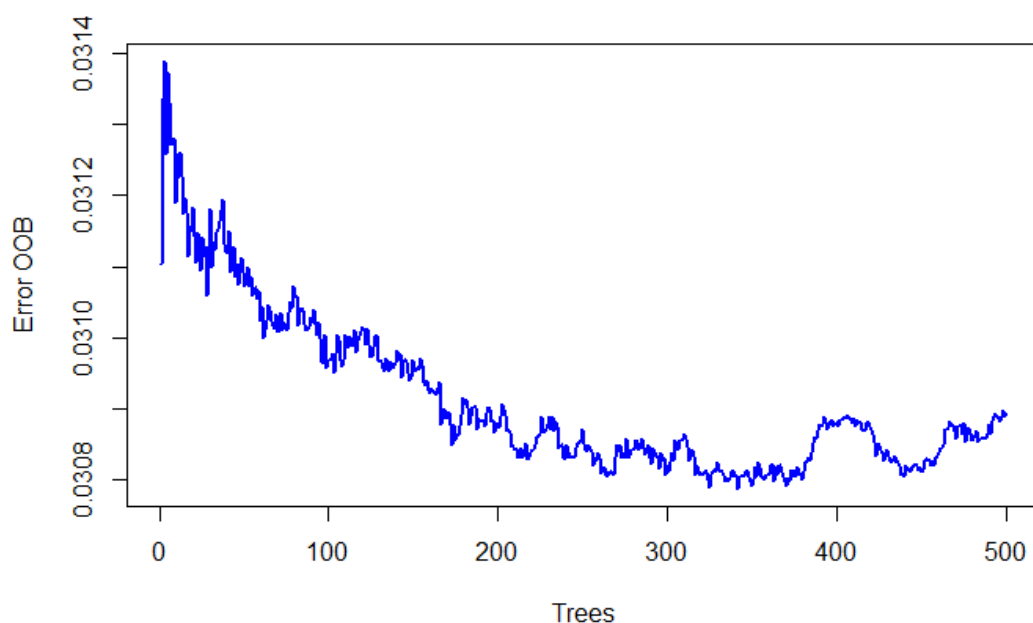


Gráfico 23– número de árvores que apresentam menor erro para Bagging

A matriz de confusão indica que o algoritmo acertou 174.333 Boletins de Ocorrência em que afirmou não ocorreu flagrante, errando apenas 1.490.

Essa técnica apresentou acurácia superior ao dos algoritmos testados anteriormente: apenas 3,13% de erro total e apresentou bom desempenho de acerto da probabilidade do registro de um BO com flagrante.

Tabela 19– Matriz de confusão para Bagging

pred_bag_f		
	0	1
0	174333	4361
1	1490	6674

3.6 Random Forest

O algoritmo Floresta Aleatória tem objetivo e formato semelhantes ao Bagging e boosting, mas seleciona as variáveis de forma diferente:

“Boosting ignores some variables completely, while the random forest does not. The candidate split-variable selection increases the chance that any single variable gets included in a random forest, while no such selection occurs with boosting” (p.612, The Elements of Statistical Learning, 2008)

Inventado pela cientista da computação Tin Kam Hoo em 1995 (Hoo, Tin, Random Decision Forests, 1995) e depois obteve contribuições de Leo Brieman e Adele Cutler (p.609, The Elements of Statistical Learning, 2008), corrigem o excessivo ajuste ao conjunto de treinamento, usando método do subespaço aleatório.

“The randomization effectively voids the effect of a variable, much like setting a coefficient to zero in a linear model. This does not measure the effect on prediction were this variable not available, because if the model was refitted without the variable, other variables could be used as surrogates” (p.612, The Elements of Statistical Learning, 2008)

E é dada pela equação:

$$\text{Regressão: } \hat{f}^B = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x})$$

Classificação: Seja $\hat{C}_b(\mathbf{x})$ a classe predita pela b -ésima árvore. Então,
 $\hat{C}^B(\mathbf{x}) = \text{voto da maioria } \left\{ \hat{C}_b(\mathbf{x}) \right\}_{b=1}^B$.

A matriz de confusão do ajuste do algoritmo de Random Forest para prever ocorrências de Flagrante de roubo de celular apresenta erro total de ligeiramente menor que o Bagging: 3,12% e praticamente o mesmo percentual de acertos de flagrante:60,48%, acertando 6.727 Boletins de ocorrência de flagrante.

Tabela 20– Matriz de confusão para Random Forest

pred_floresta_f	
0	1
0 174298	4396
1 1437	6727

O gráfico de importância de variável demonstra que autoria do BO e a Rubrica que indica a natureza jurídica, são as variáveis mais relevantes para prever ocorrências de flagrante.

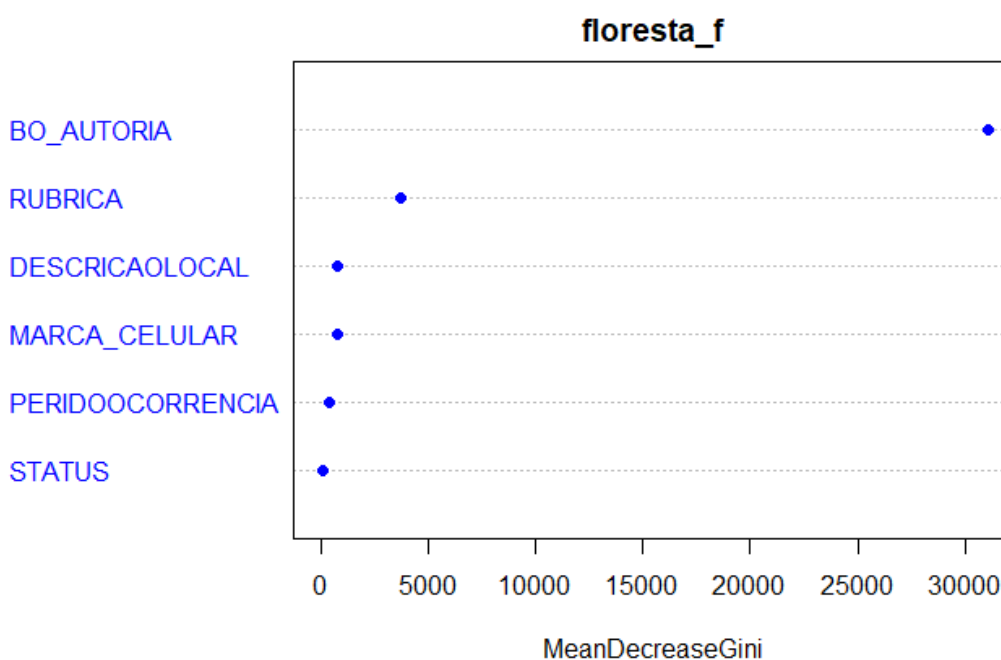


Gráfico 24– Gráfico de importância de variáveis- Random Forest

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

3.7 Support Vector Machine

Ainda em análise não supervisionada as Máquinas de Vetores de Suporte constroem um hiperplano em um espaço de dimensão infinita.

Para prever ocorrências de roubo de celular em que ocorreu flagrante foram utilizados 49.443 vetores de suporte, no kernel linear. De modo que o desempenho de erro total é de: 3,27%

Tabela 21– SVM Resultados 1

```
Call:
svm(formula = FLAGRANTE_N ~ ., data = flagrante[treino_f, ], type = "C-classification",
  kernel = "linear")

Parameters:
  SVM-Type: C-classification
  SVM-Kernel: linear
  cost: 1

Number of Support Vectors: 49443

( 24650 24793 )

Number of Classes: 2

Levels:
0 1
```

Tabela 22– SVM Resultados 2

```
[1] 0.9673067
```

3.8 Redes Neurais

Já o teste com as Redes Neurais Artificiais, utilizando o pacote Keras do RStudio, foram utilizados 128 Batches e 32 épocas, apresentando um erro total de 3,29%

Tabela 23– Redes Neurais Artificias. Resultados 1

Model: "sequential"		
Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	512
dense_1 (Dense)	(None, 32)	2080

dense_2 (Dense)	(None, 2)	66
-----------------	-----------	----

Total params: 2,658

Trainable params: 2,658

Non-trainable params: 0

Tabela 24– Redes Neurais Artificiais. Resultados

```
print(proc.time() - begin)
user system elapsed
351.92  6.73 293.75
> y_hat_net <- predict_classes(net, X_tst)
> mean(y_hat_net != y_tst)
[1] 0.03285929
```

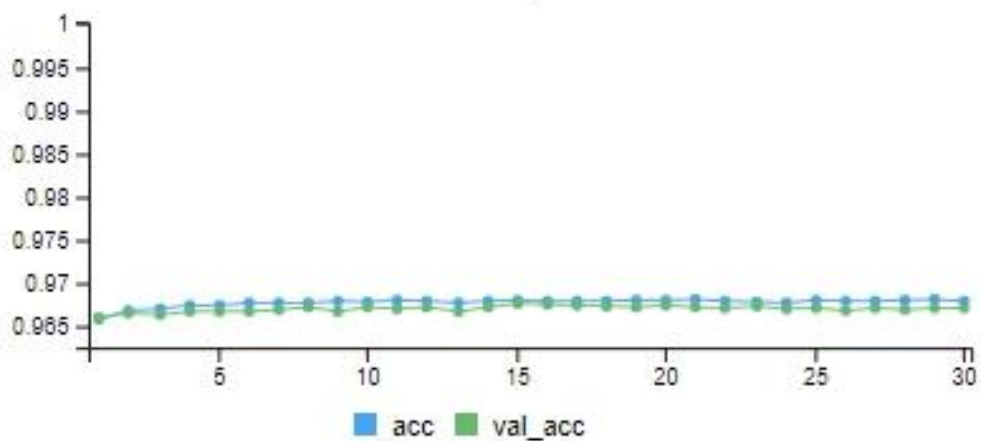
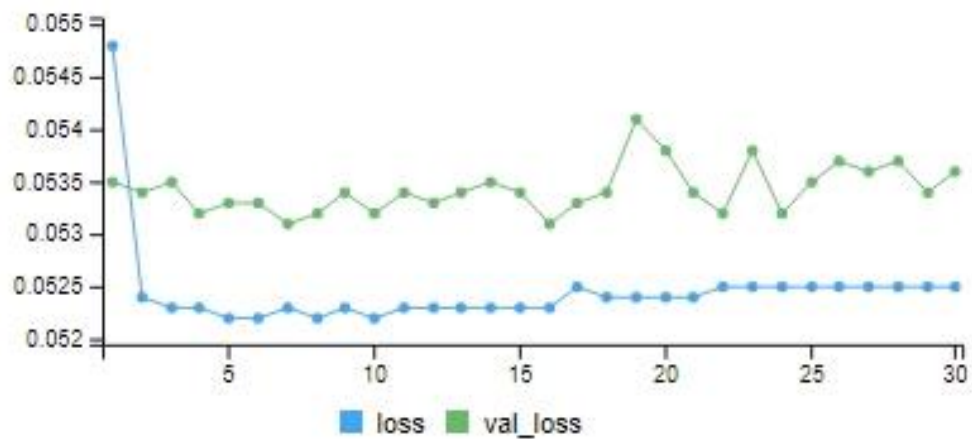


Gráfico 25 e 26– Redes Neurais Artificiais 3 e 4

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO, 2019)

5 CONCLUSÕES

A primeira parte do trabalho focada em inferência ajustou modelos lineares convencionais e espaciais. Dentre eles, há evidências de que o modelo em que as ocorrências noturnas de roubo de celular (“NOITE”) e as que o bandido estava se locomovendo com uma moto (“MOTO”) são capazes de explicar mais de 84% da variância do total de ocorrências de roubo de celular na cidade, descritas pela variável “TOTAL”. Trata-se de um bom desempenho para esse tipo de modelo, de modo que para tal resultado foram realizados todos os testes necessários de colinearidade, normalidade, análise de variância e auto correlação dos resíduos e significância estatística do modelo, além do trabalho de seleção de variáveis preditoras. Do ponto de vista da Política Pública, o objetivo de tentar prever quais variáveis e de que forma elas influenciam no número total de ocorrências de roubo de celular registradas é contribuir para mais informações referentes aos registros de Boletins de Ocorrência sobre este tipo de crime, uma vez que esses índices podem influenciar no policiamento em determinada região e de alguma forma evitar ou prevenir novos roubos.

A segunda parte do trabalho visou treinar e testar algoritmos em uma análise de Aprendizado de Máquina (Machine Learning) supervisionada, cuja a ideia era prever ocorrências de flagrante delito nos crimes de roubo de celular na cidade de São Paulo. Dentre os algoritmos treinados o Bagging, que trabalhou com 500 Árvores, apresentou melhor desempenho, sendo capaz de prever 60,48% dos casos analisados. Em termos de erro total todos eles apresentaram excelente desempenho, de modo que o melhor deles foi o Random Forest—ainda a Floresta e Bagging estejam praticamente empatados.

No entanto, ainda que tenham sido aplicadas a título de comparação com Support Vector Machines e as Redes Neurais para prever o erro total, o objetivo central da análise é prever a ocorrência de flagrante ou do Verdadeiros-Positivos.

Algoritmo	Erro total	Flagrante
Logística	3,21%	60,451%
Árvore	3,35%	59,998%
Bagging	3,13%	60,480%
RF	3,12%	60,478%
SVM	3,27%	-
Rede Neural	3,29%	-

Tabela 25– resultados: desempenho preditivo dos algoritmos testados

Enfim, é possível afirmar que os algoritmos de Random Forest e Bagging apresentam melhor desempenho para prever ocorrência de flagrante em roubo de celular na cidade de São Paulo, sendo capazes de prever cerca de 60,5% dos flagrantes deste tipo de crime na cidade.

REFERÊNCIAS

BARANDIARAN, Iñigo. The random subspace method for constructing decision forests. **IEEE Trans. Pattern Anal. Mach. Intell.**, v. 20, n. 8, p. 1-22, 1998.

BRASIL. Art. 157 DECRETO-LEI N o 2.848, DE 7 DE DEZEMBRO DE 1940. Disponível em: <http://www.planalto.gov.br/ccivil_03/Decreto-Lei/Del2848.htm>. Acesso em: 21/11/2019.

BRASIL. Secretaria da Segurança Pública do Estado de São Paulo, portal da transparência. Disponível em: <<http://www.ssp.sp.gov.br/transparenciassp/>>. Acesso em: 21/11/2019.

HASTIE, T. Tibshirani, r. and Friedman, J.(2009): The elements of statistical learning. Data mining, inference, and prediction. 2008.

HO, Tin Kam. Random decision forests. In: **Proceedings of 3rd international conference on document analysis and recognition**. IEEE, 1995. p. 278-282.