Startups: Análise Geoespacial da Relação entre o Surgimento e as Redes de Apoio Local

RESUMO

O presente artigo tem como objetivo analisar a relação entre o surgimento de *startups* e a rede de instituições de ensino de apoio local a essas *startups*, através do estudo da distribuição espacial desses fenômenos. Variáveis socioeconômico-demográficas dos municípios e áreas agregadas foram utilizadas. Para testar as hipóteses levantadas, foram utilizadas bases de *startups* do Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (Sebrae), além de bases públicas de incubadoras e aceleradoras da Associação Nacional de Entidades Promotoras de Empreendimentos Inovadores (Anprotec), de instituições de ensino do Ministério da Educação (MEC), da Receita Federal do Brasil (RFB) e do Instituto Brasileiro de Geografia e Estatística (IBGE). Foram utilizadas técnicas de análise descritiva e regressão linear múltipla, com destaque para o modelo espacial auto-regressivo (*Spatial Auto-Regressive model*). Os resultados permitem explicar 93,18% (índice de determinação, R^2) da distribuição de *startups*, a partir da disponibilidade de instituições de ensino públicas e privadas na região, da Renda Domiciliar Mensal Média, do Produto Interno Bruto e da Arrecadação de Impostos Per Capita. A explicação da existência de *startups* não é alterada significativamente considerando a influência regional por meio de regressão espacial – SAR. Isso se deve ao fato de que as variáveis independentes selecionadas para o modelo, de natureza econômica, já capturam o componente geográfico, devido a sua significativa auto-correlação espacial. Tais achados permitem endereçar melhor gestão e incentivos para a sustentação e aumento da oferta de *startups* em todo o território nacional.

Palavras-chave: *Startups*, Instituições de Ensino, Estatística Espacial

INTRODUÇÃO

A literatura sobre *startups* é ainda escassa no que se refere à investigação da taxa de surgimento e de quais fatores são capazes de predizer essa taxa de surgimento. Por mais que possamos sugerir, de maneira prescritiva, que alguns fatores podem impactar essa taxa, os estudos e a literatura acadêmica que, efetivamente, nos mostra essas relações ainda são bastante escassos.

As principais revisões nacionais sobre o tema (Silva, 2017; Nogueira e Oliveira, 2015; Minatogawa, 2013) recomendam, em suas agendas de pesquisas futuras maiores estudos que investiguem os fatores que contribuem para a taxa de sobrevivência das *startups*. De acordo com esses autores, estudos levando em considerações fatores externos às *startups* (como geolocalização, instituições de ensino da localidade e etc.) ainda são poucos.

Ainda, as *startups* se mostram importantes na economia brasileira, tendo um impacto direto na mesma (Silva, 2017). A partir de pesquisa realizada pelo Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (Sebrae) e a Associação Nacional de Entidades Promotoras de Empreendimentos Inovadores (Anprotec), sabe-se que considerando apenas as *startups* que são tendidas por incubadoras filiadas à Anprotec o faturamento entre os anos de 2012 e 2015 chegou à R$15 bilhões, enquanto os empregos diretos gerados chegaram à 53 mil (Anprotec, Sebrae 2016).

Já em relação a taxa de sobrevivência das *startups*, pesquisa conduzida por Nogueira e Oliveira (2015) nos mostra que os três principais fatores que influenciam a taxa de sobrevivência de uma *startup* são:

- Capital: de acordo com a pesquisa, quanto maior o recurso inicial da *startup*, maior a chance de falecimento da mesma. Isso porque quando não se tem capital próprio o empreendedor se vê obrigado a buscar outras fontes de recursos, em instrumentos de fomento e também com investidores.

- Relação societária: quanto mais sócios, maior a chance de problemas de relacionamento entre esses sócios, e consequentemente maior a probabilidade de fechamento.
- Localização: Dados da pesquisa mostram que *startups* residentes em incubadoras, parques tecnológicos, aceleradoras e etc., reduzem em 3,45 vezes a chance de quebrar quando comparadas com *startups* instaladas em locais próprio e sem interação.

Este último item também corrobora para o que analisaremos no presente trabalho: a relação desses instrumentos de apoio com o surgimento e a taxa de sobrevivência das *startups*.

Pretendemos validar as seguintes hipóteses:

- Quanto mais instrumentos de apoio local (como parques tecnológicos, incubadoras e etc.), maior a quantidade de *startups* para esta mesma localidade.
- Regiões com variáveis econômicas mais desenvolvidas possuem maior densidade de *startups* formalizadas.

Dessa forma, considerando as hipóteses acima, pretendemos responder à seguinte questão: *Há influência entre a existência de instituições de ensino e outros instrumentos locais de apoio com a quantidade de startups existentes em determinada localidade?*

OBJETIVOS DO ESTUDO

Temos como objetivo geral relacionar o surgimento de *startups* com a existência de instituições de ensino e outros instrumentos locais de apoio, como incubadoras e parques tecnológicos. Os objetivos específicos de nossa pesquisa foram:

- Avaliar os efeitos preditivos entre as variáveis surgimento de *startups* (*startups* novas – formalizadas) e existência de instituições de ensino.
- Avaliar os efeitos preditivos entre as variáveis econômicas e surgimento de *startups* (*startups* novas – formalizadas).

REFERENCIAL TEÓRICO

STARTUPS

Quando se pesquisa sobre *startups*, percebe-se que as mesmas possuem a agilidade como uma de suas características chaves, além de apresentarem um rápido e elevado crescimento, e de saberem lidar melhor em cenários de risco e incertezas (Weiblen e Chesbrough, 2015).

Porém, temos um desafio em relação à definição de *startups* (Minatogawa, 2013), já que diferentes autores conceituam esse tipo de empresa de forma diferente entre si. Não há um único conceito (academicamente ou mercadologicamente falando). Dessa forma, apresentaremos no presente trabalho alguns conceitos mais comuns no meio acadêmico, seguido do conceito escolhido para utilizarmos no presente trabalho.

Segundo Ries (2012), um dos acadêmicos mais reconhecidos na área de inovação em relação aos seus estudos em relação à temática, uma *startup* pode ser definida como uma "instituição humana, desenhada para criar um novo produto ou serviço, em condições de extrema incerteza".

Já Dullius e Schaeffer (2016) nos diz que *startups* são "empresas nascentes de base tecnológica, que possuem na inovação tecnológica disruptiva os fundamentos de sua estratégia competitiva e que passam a competir com grandes empresas previamente estabelecidas em termos de inovação".

Por outro lado, Torres (2017) nos traz que *startups* estão focadas em ter um produto ou serviço que solucione um problema, mas não focam em pessoas. Essa conceituação corrobora uma frase bastante famosa no meio de *startups*, do Uri Levine, cofundador do Waze, que diz "apaixone-se pelo problema, e não pela solução".

Assim, tendo em vista a quantidade de diferentes conceitos para o que vem a ser uma *startup*, para o presente trabalho, consideramos a conceituação utilizada pelo Sebrae (2017), por ser uma organização de relevante representatividade no cenário brasileiro de *startups*: uma *startup* é um grupo de pessoas à procura de um modelo de negócios repetível e escalável, trabalhando em condições de extrema incerteza.

ACELERADORAS E INCUBADORAS

As primeiras aceleradoras surgiram no Brasil entre 2010 e 2011 (Aguilhar, 2014), são elas:

- Aceleradora (de São Paulo) e
- 21212 (do Rio de Janeiro).

Isso faz com que a literatura sobre esse assunto seja ainda mais recente, e consequentemente que também tenhamos pouco arcabouço teórico consolidado sobre o assunto no Brasil. Porém, enxergamos isso como uma oportunidade, já que temos um campo fértil para estudos relacionados a essa temática.

Quando se compara as incubadoras e as aceleradoras percebe-se que as primeiras geralmente são ligadas a instituições públicas e possuem um foco de contribuir com que *startups* acessem subsídios, além de não aportarem recursos diretamente na *startup*. As aceleradoras, por outro lado, costumam ser iniciativas privadas, com fins lucrativos, que conectam *startups* à mentores e ao mercado, e geralmente fazem um aporte financeiro inicial na *startup* (Ribeiro, Plonski e Ortega, 2015).

O que percebemos é que, cada vez mais, instituições de apoio têm surgido no país, tanto as ligadas ao governo e instituições públicas, como também as privadas, com fins lucrativos. De acordo com a Anprotec e Sebrae (2016) "a oferta de ambientes propícios à criação e ao crescimento de negócios e soluções inovadoras tem se mostrado, ao longo dos anos, uma ferramenta importante de impulso ao desenvolvimento econômico, tecnológico e social".

PROCEDIMENTOS METODOLÓGICOS

Utilizando-se o critério de classificação proposto por Vergara (2005), pode-se classificar uma pesquisa de acordo com dois aspectos, a saber: quanto aos fins e quanto aos meios de investigação.

Quanto aos fins, a pesquisa é do tipo exploratória e descritiva, pois obteve e compreendeu dados acerca de um determinado fenômeno até então pouco explorado na literatura científica sobre comportamento organizacional.

Além disso, a pesquisa é, também, de natureza inferencial, pois procurou examinar relações entre variáveis, de forma a procurar quais fatores influenciam para a ocorrência de um certo fenômeno. Note-se que foi utilizada abordagem quantitativa para análise dos resultados obtidos.

A partir de dados coletados de diversas fontes, utilizaremos como variável resposta a existência de *startups* em determinada região intermediária para avaliar a relação com a existência de instituições de ensino e outros instrumentos locais de apoio, como incubadoras e parques tecnológicos. Além disso, utilizaremos técnicas de estatística descritiva e geoespacial para aprofundar a análise dessa relação.

No próximo item remos discorrer sobre a escolha de regiões intermediárias para, em seguida, apresentar dos modelos de regressão utilizados no trabalho e, por fim, apresentar a metodologia CRISP-DM, utilizada no processo de entendimento do problema de pesquisa e mineração de dados para atingir os objetivos propostos.

RECORTE GEOGRÁFICO DA PESQUISA: REGIÕES INTERMEDIÁRIAS

Devido à necessidade de atualização do quadro regional anterior (Mesorregiões e Microrregiões Geográficas) produzido na década de 1980, o recorte das Regiões Geográficas Imediatas (RGI) e Intermediárias (RGINT) de 2017 foi estabelecido para incorporar as mudanças ocorridas no Brasil ao longo das últimas três décadas. (IBGE, 2017)

O IBGE baseou a nova proposta de regionalização em outros estudos do próprio órgão. Os estudos foram:

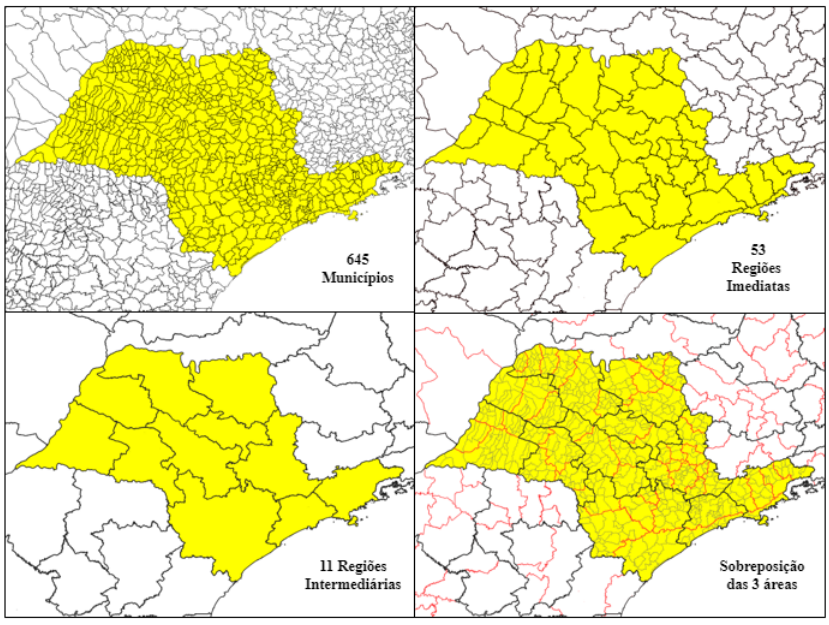
- Regiões de Influência das Cidades - Regic 2007
- Divisão Urbano-Regional
- Gestão do Território 2014
- Logística dos Transportes do Brasil 2014
- Arranjos Populacionais e Concentrações Urbanas do Brasil

As Regiões Geográficas Imediatas têm na rede urbana o seu principal elemento de referência. Essas regiões são estruturas a partir de centros urbanos próximos para a satisfação das necessidades imediatas das populações, tais como: compras de bens de consumo duráveis e não duráveis; busca de trabalho; procura por serviços de saúde e educação; e prestação de serviços públicos, como postos de atendimento do Instituto Nacional do Seguro Social - INSS, do Ministério do Trabalho e de serviços judiciários, entre outros (IBGE, 2017).

As Regiões Geográficas Intermediárias correspondem a uma escala intermediária entre as Unidades da Federação e as Regiões Geográficas Imediatas. As Regiões Geográficas Intermediárias foram criadas respeitando um número mínimo de duas Regiões Geográficas Imediatas para cada Região Geográfica Intermediária. Foi definido que cada UF deve conter, no mínimo, duas Regiões Geográficas Intermediárias. (IBGE, 2017)

A Figura 1 mostra a diferença entre as áreas de municípios, regiões imediatas e regiões intermediárias, no exemplo do Estado de São Paulo. A sobreposição das 3 subdivisões (apresentada no mapa de baixo à direita) sugere a escolha das regiões intermediárias para este estudo.

Figura 1. Comparação entre áreas de abrangência do IBGE



Fonte: IBGE (2010, 2017)

Como a distribuição geográfica das observações estudadas se mostraram concentradas em um número reduzido de municípios, regressões por municípios ou mesmo por regiões imediatas trouxeram grande quantidade de valores em branco. Considerando esse fator, optamos, neste estudo, por nos basearmos em regiões intermediárias com o objetivo de obter melhores resultados de análise de peso geográfico. A seguir, abordaremos os conceitos dos Modelos de Regressão utilizados em nossa análise.

MODELOS DE REGRESSÃO

Foram utilizadas como técnicas de regressão a regressão tradicional, e uma técnica geoestatística: *Spatial Auto-Regression* (SAR). Todas são técnicas exploratórias para análise de dados.

A Regressão Linear Multivariada é uma técnica estatística que permite estimar a probabilidade de ocorrência de um evento a partir de um conjunto de informações que caracterizam esse evento (Sicsú, 2010).

Já a *Spatial Auto-Regression* (SAR) remonta aos anos 1950 é tradicionalmente adotada em estatística espacial (De Smith, Goodchild e Longley, 2007). Utiliza uma matriz de vizinhança (W), associada ao próprio termo dependente y na estimação de y. Por isso, é denominada uma técnica auto-regressiva (Francisco et al., 2008).

As Equações 1 e 2 a seguir mostram como o termo auto-regressivo é incorporado, e a Figura 1 exemplifica a estrutura da matriz de vizinhança:

Equação 1: Regressão Linear Multivariada : $\boldsymbol{y = \beta X + \varepsilon}$

Equação 2: Regressão Espacial (SAR) : $\boldsymbol{y = \rho W y + \beta X + \varepsilon}$

em que $\boldsymbol{y}$ é a variável dependente; $\boldsymbol{X}$ é o vetor de variáveis independentes, $\boldsymbol{\beta}$ é o vetor de coeficientes e $\boldsymbol{\varepsilon}$ é o erro da estimação (supostamente normal).

Figura 2. Exemplo de Matriz de Vizinhança



Fonte: Francisco et al. (2008)

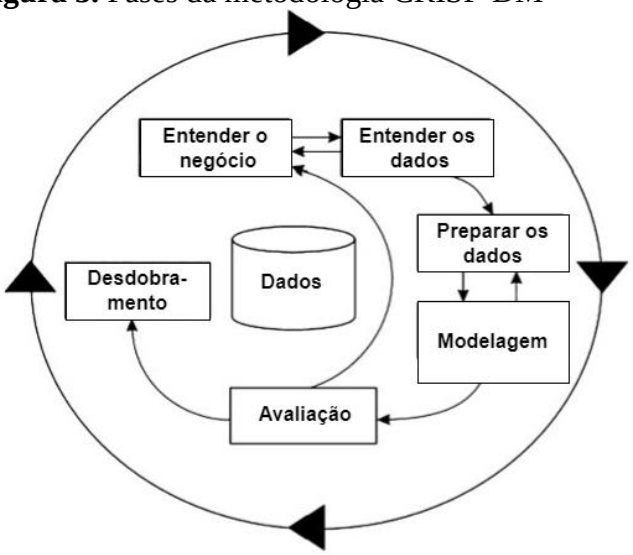
METODOLOGIA CRISP-DM

A metodologia *Cross - Industry Standard Process of Data Mining* ou CRISP-DM foi criada durante a década de 90 por uma necessidade de compilar tentativas anteriores de definir metodologias para padronizar o processo de mineração de dados com foco eu auxiliar a tomada de decisão. Wirth & Hipp (2000) citam que os autores que subsidiaram, com seus estudos, a criação dessa metodologia como sendo: Reinartz e Wirth (1995), Adriaans e Zantinge (1996), Brachman e Anand (1996) e Fayyad et al. (1996).

A estrutura do CRISP-DM tem como objetivo auxiliar os pesquisadores desde o planejamento até a execução da mineração de dados, passando pela especificação do processo da descoberta do conhecimento até a apresentação dos resultados alcançados (Pádua e Sousa, 2018).

A metodologia CRISP-DM é composta por 6 fases, organizadas de maneira cíclica, cujo fluxo possibilita ir e voltar entre as essas fases, que compreendem: a) *Business Understanding* (Entendimento do Negócio); b) *Data Understanding* (Entendimento dos Dados); c) *Data Preparation* (Preparação dos Dados); d) *Modeling* (Modelagem); e) *Evaluation* (Avaliação); f) *Deployment* (Desdobramento do Modelo) (Chapman et al., 2000). A Figura 3 representa a sequência de cada uma das fases.

Figura 3. Fases da metodologia CRISP-DM



Fonte: Adaptado de Wirth & Hipp (2000)

Em nosso estudo, adotamos essa metodologia de análise de dados com foco no entendimento dos negócios para nortear todo o processo de tratamento e modelagem dos dados. O entendimento do negócio foi subsidiado pela revisão da literatura, bem como pela expertise dos autores, a análise exploratória realizada serviu para entendimento dos dados e subsidiou, junto do entendimento dos negócios, a preparação e a modelagem dos dados.
Seguindo o ciclo do CRISP-DM, a avaliação do modelo nos deu subsídios para compreensão maior do negócio, permitindo as análises realizadas, além da proposição de aprimoramento do modelo apresentado, bem como a apresentação de sugestões de estudos futuros sobre o negócio.

ANÁLISE DOS RESULTADOS

A variável resposta, existência de *startups* (T_STP), foi construída a partir de dados coletados do Sebrae e do Cadastro de Pessoas Jurídicas da Receita Federal de Brasil. Esses dados foram posteriormente sumarizados em regiões intermediárias.
Conforme abordamos anteriormente, a escolha por regiões intermediárias se baseou no fato de que, se optássemos por outro tipo de agrupamento (região imediata, município, setor censitário) teríamos muitos *missing values* que, hipoteticamente, poderiam influenciar o entendimento do problema.
Com base no conhecimento do negócio, as variáveis que propomos para explicar a existência de *startups* foram: instituições de apoio, instituições de ensino público, instituições de ensino privado. Além delas, foram incluídos dados demográficos do IBGE, como PIB per capita, imposto per capita, população economicamente ativa, renda per capita, média do IDH, renda domiciliar mensal média, total de domicílios, valor de receita por instituição de ensino per capita, percentual de *startups* ativas, percentual de *startups* baixadas e média do tempo de vida das *startups* na região.

PREPARAÇÃO DOS DADOS

Para realizar o trabalho, utilizamos bases diversas, elencadas na Tabela 1. Utilizamos as ferramentas R, Python, o software *desktop GIS* denominado QGIS 3.14 e a API do Google Maps/Earth nesta fase de preparação.

Tabela 1. Bases de dados utilizadas no estudo

Base	Ano	Origem	Registros	Atributos	Geocodificados
Startups Brasil	2020	Sebrae Nacional	12626	3	5967
Aceleradoras, Incubadoras e Instituições de Apoio	2020	Anprotec	359	10	359
Instituições de Ensino	2018	INEP	3118	21	2448
Censo da Educação Superior	2018	MEC	2448	48	2050
Base CNPJ do Brasil	2020	RFB / qsacnpj	43101755	40	5731
Dados Geográficos do Brasil - Censo	2010	IBGE	510	11	5570
Dados Geográficos do Brasil - regiões intermediárias	2017	IBGE	133	2	133
API *Geocoding*	2020	Google Maps/Earth	N/A	5	5731

Fonte: Os autores

A primeira etapa foi realizada no software R. Para realizar a geocodificação, o objetivo foi usar uma base confiável de origem para o CEP, uma vez que a base de *startups* proveniente do Sebrae Nacional foi preenchida manualmente e sem verificação. Seguindo este critério, foi adotada a base da Receita Federal do Brasil - RFB para enriquecimento deste dado.

Experienciamos um grande desafio no tratamento dessa base, devido a capacidade de processamento das máquinas que tínhamos disponíveis. Tínhamos à disposição uma base de dados com dezenas de milhões de registros, o que demandou maior capacidade computacional e, em consequência, mais tempo de processamento. Isso motivou a alteração da abordagem de tratamento dos dados para ganhos de performance.

Assim, adotamos uma abordagem via R + sqlLite, na qual as bases que seriam enriquecidas foram inseridas neste banco para a realização de *joins*, consumindo menos recursos. Para a ingestão neste banco foram considerados os registros de CNPJ com 14 dígitos e preenchidos com zeros à esquerda dada a característica da base para a realização do *join*.

A mesma tratativa foi realizada com as instituições de ensino, após a base ter sido enriquecida com os CNPJ retirados da base do INEP. Os CEP que foram retornados das *startups* e instituições de ensino foram submetidos à API de *geocoding* do Google, gerando uma tabela de coordenadas por CEP atribuídos aos registros completos.

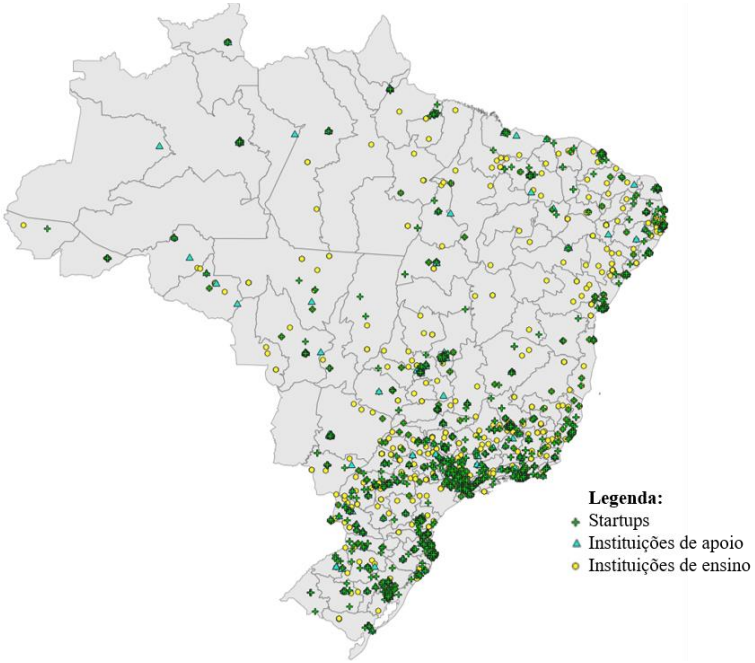
Para o cálculo do tempo de vida, utilizamos a diferença entre a data de abertura das empresas abertas do ano 2000 em diante até a data de 28/02/2020 (data da extração dos dados na RFB, caso a empresa estivesse baixada o cálculo foi realizado até a data de baixa. A situação cadastral da empresa foi tratada em formato de *dummy* de maneira binária (ativa/baixada).

Para manipular os dados da Anprotec, utilizamos Python. O site da associação, na data de nosso acesso, apresentava um mapa com as instituições de apoio. Realizamos a raspagem destes dados que estavam em um *javascript* que era o responsável por plotar estes dados no mapa.

Por fim, nessa etapa de preparação dos dados, utilizamos o QGIS para sumarização dos dados. Com os registros geocodificados, todos os dados foram sumarizados por região intermediária via *join* tabular. Os dados do IBGE também foram sumarizados pelo mesmo tema e agregados

juntamente com as instituições de ensino e apoio. Criamos, então, campos calculados com as taxas de *startups*, impostos e PIB per capita e a criação de mapas temáticos para verificação da distribuição das variáveis.

Figura 4. Geocodificação – Distribuição Espacial das IES, Instituições de Apoio e *Startups*



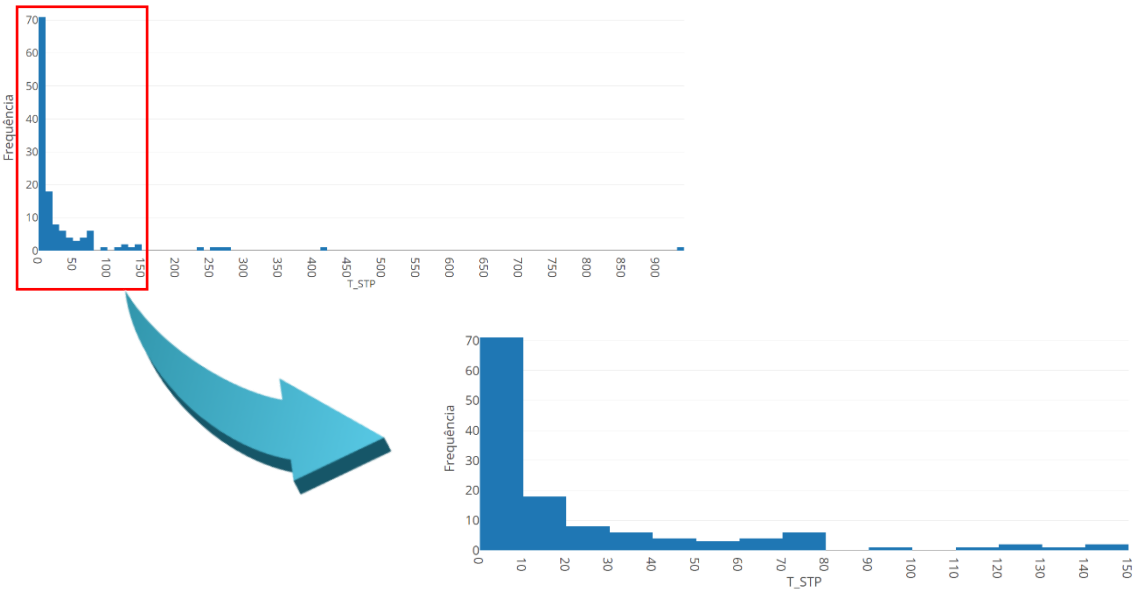
Fonte: Os autores, a partir do uso da ferramenta QGIS 3.14

Após o processo de geocodificação representado na Figura 4, identificamos que, das 133 regiões intermediárias do Brasil, 18 não possuem *startups*, mas contam com algum tipo de entidade de apoio a *startups* ou instituições de ensino. Ainda, verificamos que 7 regiões intermediárias, além de não terem *startups*, não contam com nenhuma entidade de apoio e instituição de ensino, comprovando a pertinência na escolha desse recorte regional.
A seguir, continuaremos a abordar o processo de entendimento dos dados, a partir de sua exploração e preparação dos dados para modelagem, demonstrando o processo de seleção de variáveis, incluindo dados econômicos aos pontos geocodificados para compor o modelo.

EXPLORAÇÃO DE DADOS

Em relação à distribuição da quantidade de *startups*, verificamos assimetria devido a quantidade de *startups* estar concentrada em determinadas regiões e não possuírem uma distribuição uniforme com valores distantes da média.
A seguir, apresentamos no Gráfico 1 o histograma de distribuição de *startups*. Os pontos isolados à direita do histograma representam as regiões intermediárias com as maiores quantidades de *startups*. São Paulo, Rio de Janeiro, Porto Alegre, Curitiba, Belo Horizonte e Florianópolis com, respectivamente, totais de 931, 418, 275, 268, 258, 232.
O gráfico apresenta ainda o detalhamento da primeira barra de frequência do Gráfico 1, excluindo-se as 5 regiões que possuem maior quantidade de *startups*. Nele, verificamos 71 regiões intermediárias que possuem até 9 *startups*. 25 regiões não possuem *startups* em seu território.

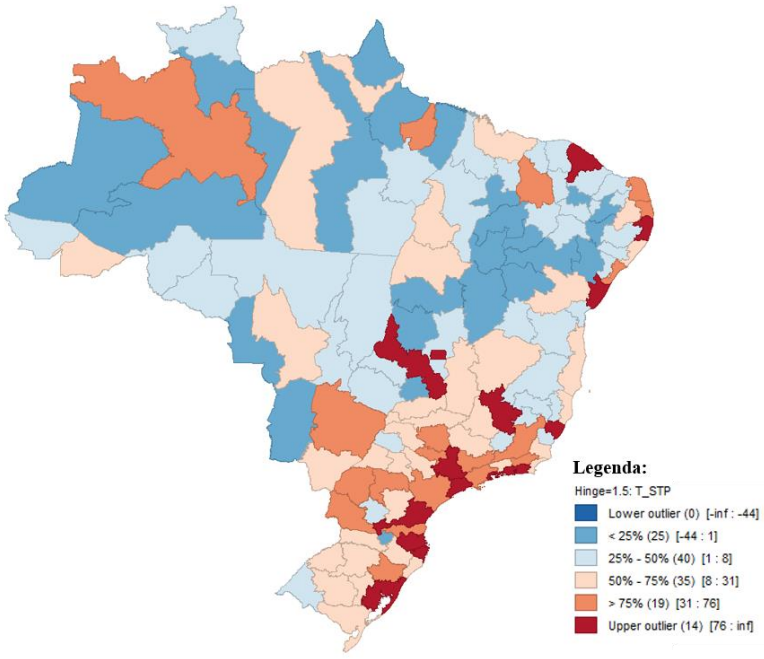
Gráfico 1. Histograma de distribuição de *Startups*



Fonte: Os autores, a partir do uso da ferramenta R 3.6.2
Nota: Histograma de baixo destaca a distribuição sem as 5 regiões com maior quantidade de *startups*.

Com os dados agregados pelo QGIS, foi criada a matriz de vizinhança utilizando critério de contiguidade Queen de ordem 1, bem como as análises espaciais e as regressões propostas (múltipla e espacial).
A Figura 5 apresenta o *box map* (mapa temático com a variável categorizada a partir dos quartis) da quantidade de startups por Região Intermediária. Percebe-se que as regiões Sul e Sudeste são as regiões que concentram maior quantidade de *startups*.

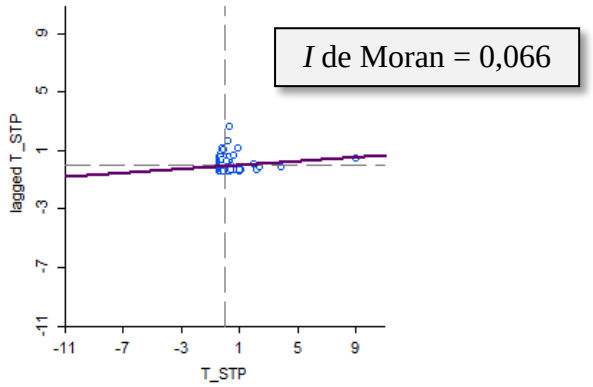
Figura 5. *Box Map* quantidade de startups T_STP por região intermediária



Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: n = 133

Calculamos a auto-correlação espacial (Índice *I* de Moran) da variável resposta, T_STP, a partir de uma matriz Queen de ordem 1, com o propósito de buscar indícios a respeito da validade da hipótese de que existe influência da região geográfica, contendo instituições de ensino e entidades de apoio na existência de *startups*. Não há correlação espacial significativa na variável resposta, conforme verificamos no Gráfico 2:

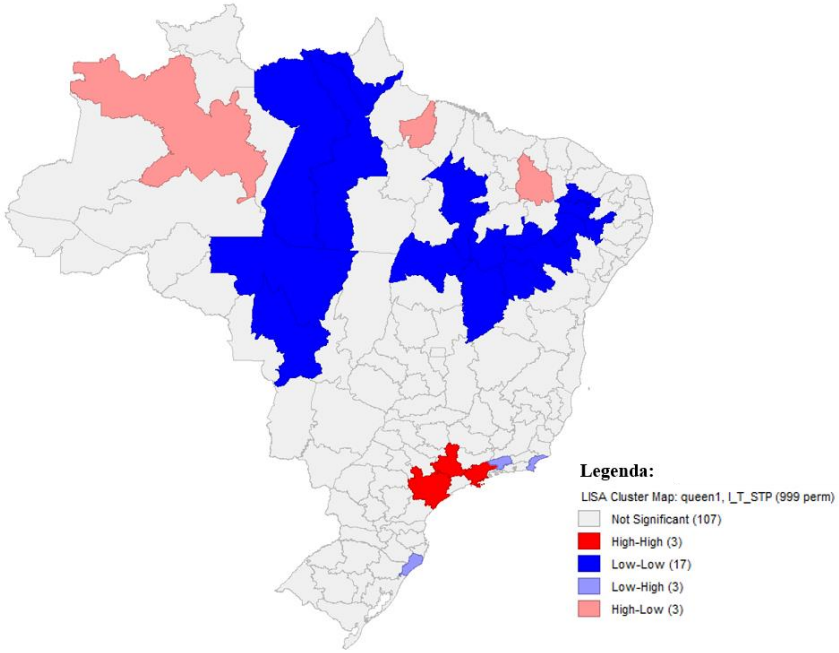
Gráfico 2. *I* de Moran do Total de Startups (T_STP)



Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: *n* = 133

Ao analisar a Figura 6, percebemos que as regiões intermediárias de Sorocaba, Campinas e São José dos Campos são high-high, regiões com alta quantidade de *startups*, "avizinhadas" por regiões com altas quantidades de *startups*, perfazendo um total de 257.

Figura 6. Mapa de *Clusters Espaciais* - Total de Startups (T_STP)



Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: *n* = 133

A partir da análise dos bancos de dados dispostos no Quadro 1 e escolha preliminar de variáveis, com base no conhecimento do negócio, nosso *dataset* contemplou as seguintes informações:

Quadro 1. Dados selecionados para avaliação de utilização no modelo

Campo	**Descrição**
RGINT	Código região imediata
NOME_RGINT	Descrição região Imediata
PCT_PIB	PIB per capita
T_DOM	Total domicílios 2010
PCT_RENDA	Renda per capita 2010
AVG_IDH_10	IDH 2010
PEA_10	População economicamente ativa 2010
AVG_RENDOM	Renda domiciliar mensal média 2010
T_APOIO	Total instituições de apoio
T_PUBLICA	Total de instituições de ensino privada
T_PRIVADA	Total de instituições de ensino pública
T_*STARTUP*	Total de *startups*
T_STP_BX	Total de *startups* baixadas
T_STP_AT	Total de *startups* ativas
AVG_TTL	Média de tempo de vida empresa na região
PCT_IMP	Imposto per capta
TX_MORTE	% *startups* ativas
TX_ATIVA	% *startups* baixadas

Fonte: Os autores

A partir desses dados, iniciamos a etapa de modelagem. Discorreremos, agora, sobre esse processo, que foi realizado utilizado o software GeoDa 1.14. Primeiramente, utilizamos a Regressão Múltipla para determinar a existência de *startups* seria influenciada pela existência de instituições de ensino e entidades de apoio, além dos dados econômicos das regiões intermediárias, que foram selecionadas previamente pelo entendimento do negócio.
Após a seleção de variáveis com base no conhecimento do negócio, utilizamos o valor-P inferior a 0,05 como critério de seleção de variáveis para compor o modelo. Para isso, utilizamos o pacote do R SignifReg, função *stepwise* para avaliação das variáveis. As variáveis utilizadas nessa função foram as seguintes:

Quadro 2. Dados selecionados para os Modelos de Regressão e significância estatística no Modelo Conjunto

Campo	**Descrição**	*Valor-P*
PCT_PIB	PIB per capita	0,34528
T_DOM	Total domicílios 2010	0,10204
PCT_RENDA	Renda per capita 2010	0,00000
AVG_IDH_10	IDH 2010	0,00000
PEA_10	População economicamente ativa 2010	0,00190
AVG_RENDOM	Renda domiciliar mensal média 2010	0,04507
T_APOIO	Total instituições de apoio	0,50257
T_PUBLICA	Total de instituições de ensino privada	0,00472
T_PRIVADA	Total de instituições de ensino pública	0,60658
T_STP_AT	Total de *startups* ativas	0,00000
T_STP_BX	Total de *startups* baixadas	0,00000
AVG_TTL	Média de tempo de vida empresa na região	0,64075
PCT_IMP	Imposto per capita	0,97620

Fonte: Os autores, a partir do uso da ferramenta R 3.6.2

Após o processo de *stepwise*, as seguintes variáveis foram consideradas para o modelo, todas elas com valor-P inferior a 0,05. Para evitar o sobre ajuste em função da multicolinariedade tratamos a correlação entre as variáveis a partir do cálculo do VIF (*Variance Influence Factor*), excluindo as variáveis com valor maior que 10. O modelo obtido trouxe as variáveis preditoras listadas na Tabela 2.

Tabela 2. Dados selecionados para rodada do *stepwise*

Código	Estimativa	Erro Padrão	*t-value*	Pr >|*t*|	VIF
(Intercept)	-3,72E+04	7,91E+03	-4,700	6,70e-06	
T_DOM	4,76E-02	1,04E-02	4,588	1,07e-05	10,62 (retirada)
T_PRIVADA	7,63E+02	1,82E+02	4,196	5,08e-05	6,94
AVG_RENDOM	2,76E+01	6,58E+00	4,198	5,05e-05	5,51
T_PUBLICA	5,35E+03	9,21E+02	5,810	4,80e-08	6,71
PCT_PIB	-1,80E+00	4,55E-01	-3,959	0,000125	6,88
PCT_IMP	8,55E+03	2,70E+03	3,167	0,001931	6,71

Fonte: Os autores, a partir do uso da ferramenta R 3.6.2
Nota: *n* = 133

A partir daí, calculamos a Regressão Linear Multivariada, na qual verificamos alto R^2 (93,18%) conforme Figura 7:

Figura 7. Regressão Linear Multivariada

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  rgint_18var
Dependent Variable  :       T_STP  Number of Observations:  133
Mean dependent var  :     38.5338  Number of Variables   :    6
S.D. dependent var  :     99.0928  Degrees of Freedom    :  127

R-squared           :    0.931803  F-statistic           :     347.051
Adjusted R-squared  :    0.929118  Prob(F-statistic)     :           0
Sum squared residual:     89063.5  Log likelihood        :    -621.418
Sigma-square        :     701.288  Akaike info criterion :     1254.84
S.E. of regression  :     26.4818  Schwarz criterion     :     1272.18
Sigma-square ML     :     669.651
S.E of regression ML:     25.8776

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      -29.5971        8.32386       -3.5557     0.00053
         T_PRIVADA       1.20709       0.165611        7.2887     0.00000
        AVG_RENDOM      0.027802     0.00708337        3.92497     0.00014
         T_PUBLICA       7.88699       0.793386        9.94092     0.00000
           PCT_PIB   -0.00217188    0.000481278       -4.51273     0.00001
           PCT_IMP        12.346         2.7628        4.46867     0.00002
------------------------------------------------------------------------------
```

Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: *n* = 133

Para verificar se que havia influência regional na existência de *startups*, promovemos o modelo obtido a partir de variáveis a uma regressão espacial múltipla - SAR. Utilizamos o modelo a partir de uma matriz de vizinhança Queen de Ordem 1.
A partir do coeficiente de determinação obtido, pudemos notar que a inclusão do peso das regiões intermediárias vizinhas na existência de *startups* não se mostrou relevante,

considerando que o índice se manteve praticamente o mesmo: 93,18% na Regressão Múltipla para 93,22% na Regressão Espacial.

Figura 8. *Spatial Auto-Regressive* (SAR) *model*

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : rgint_18var
Spatial Weight      : queen1
Dependent Variable  :       T_STP  Number of Observations:  133
Mean dependent var  :     38.5338  Number of Variables   :    7
S.D. dependent var  :     99.0928  Degrees of Freedom    :  126
Lag coeff.   (Rho)  :  -0.0499561

R-squared           :    0.932255  Log likelihood        :    -621.009
Sq. Correlation     : -            Akaike info criterion :     1256.02
Sigma-square        :     665.216  Schwarz criterion     :     1276.25
S.E of regression   :     25.7918

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
          W_T_STP    -0.0499561       0.055508       -0.89998     0.36813
         CONSTANT         -28.51        8.19294       -3.47982     0.00050
        T_PRIVADA         1.2099       0.161501        7.49162     0.00000
       AVG_RENDOM     0.0269582     0.00696397        3.87109     0.00011
        T_PUBLICA         7.8549       0.773348         10.157     0.00000
          PCT_PIB   -0.00212811    0.000473236       -4.49693     0.00001
          PCT_IMP       12.8479        2.71936         4.7246     0.00000
------------------------------------------------------------------------------
```

não significante

Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: *n* = 133

Em relação às variáveis selecionadas, verificamos também a baixa correlação espacial das variáveis Total de Instituições Públicas, Total de Instituições Privadas e Total de Domicílios. Variáveis como Renda Domiciliar, impostos per capita e PIB per capita são auto correlacionadas espacialmente. São variáveis econômicas que são conhecidas por tal característica. A seguir, apresentamos os índices de *I* de Moran e o *LISA Map* de cada variável utilizada no nosso estudo nos mapas da Figura 9. Identificamos, portanto, relação positiva entre quantidade de *startups* e existência de instituições de apoio e de ensino na região. Foi possível identificar, na análise exploratória, a concentração de *startups* por região intermediária, bem como a natureza das atividades. Porém, não foi possível identificar melhorias no modelo utilizando SAR. A probabilidade do coeficiente espacial do total de startups na região não se mostrou significante, 36,81%, conforme Figura 8. Isso se deve ao fato de que as variáveis independentes selecionadas para o modelo, de natureza econômica, já capturam o componente geográfico. Assim, ao utilizarmos a Regressão Espacial Múltipla, não conseguimos melhora significativa do modelo.

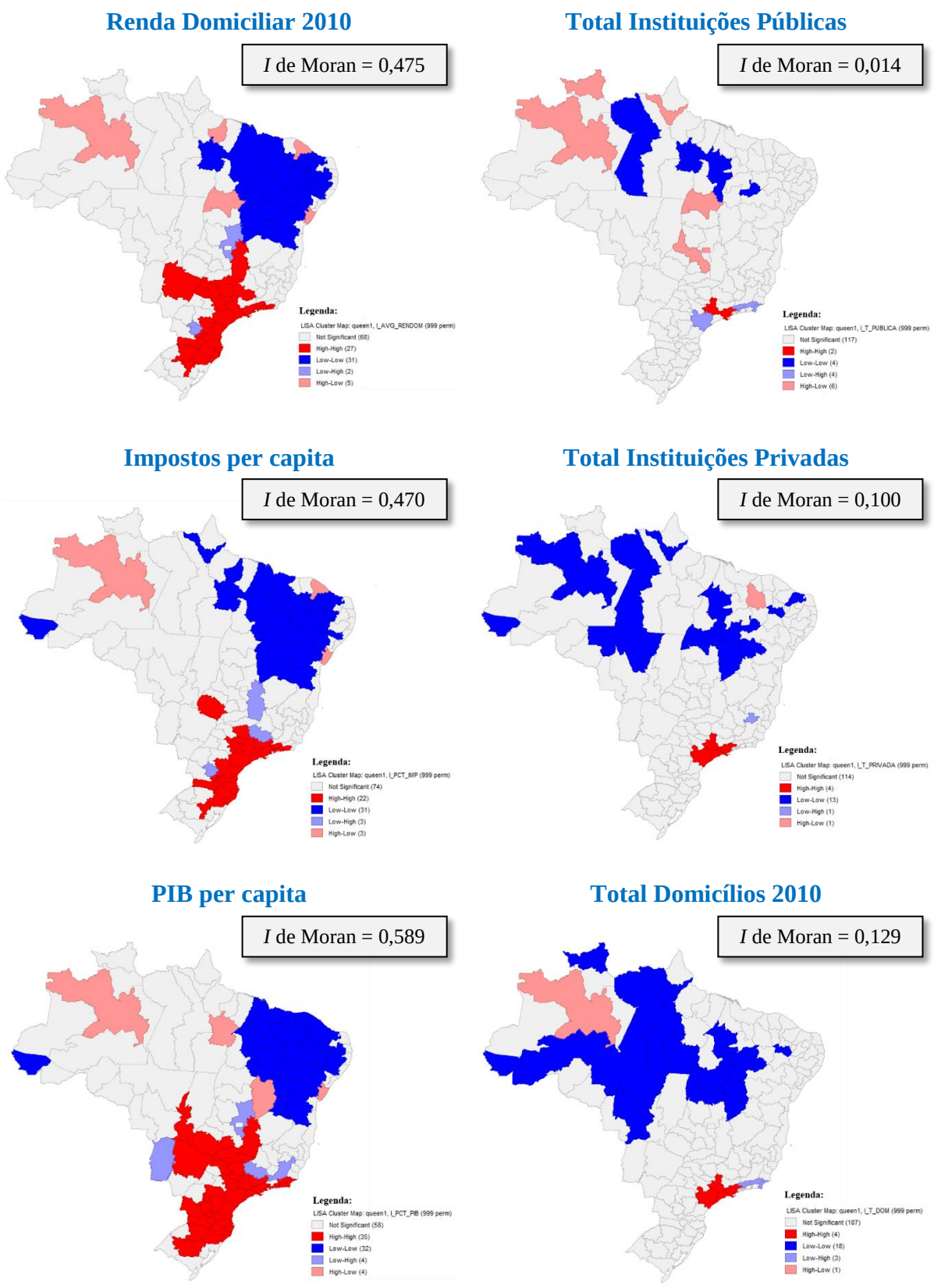
CONCLUSÕES, RECOMENDAÇÕES E SUGESTÕES PARA ESTUDOS FUTUROS

Consideramos nosso estudo de extrema relevância para a exploração do tema sob o aspecto acadêmico, considerando o número reduzido de trabalhos relevantes nessa área e, também, pela importância da temática de *startups* para o desenvolvimento socioeconômico do país.

Atingimos o objetivo geral de relacionar o surgimento de *startups* com a existência de instituições de ensino e outros instrumentos locais de apoio, como incubadoras e parques tecnológicos. O modelo final de Regressão Linear Múltipla apresentou explicação de 93%.

Em relação aos objetivos específicos, foi possível avaliar os efeitos preditivos entre as variáveis surgimento de *startups* (*startups* novas – formalizadas) e existência de instituições de ensino, bem como os efeitos preditivos entre as variáveis econômicas e surgimento de *startups* (*startups* novas – formalizadas).

Figura 9. *LISA Map* (Mapa de Clusters Espaciais) das Variáveis Preditoras



Fonte: Os autores, a partir do uso da ferramenta GeoDa 1.14
Nota: n = 133

Verificamos que as variáveis independentes selecionadas para o modelo, de natureza econômica, já capturam o componente geográfico, mais fortemente do que a variável dependente deste estudo. Assim, ao utilizarmos a Regressão Espacial Múltipla, não conseguimos melhora significativa do modelo.

Identificamos como limitação do estudo a diferença de temporalidade das Bases de dados utilizadas. Os dados mais atualizados relacionados ao IBGE são de 2010 e 2017, enquanto os demais dados disponíveis são de 2020 (Receita Federal, Sebrae e Anprotec), 2018 (Inep e MEC). Inicialmente, a nossa proposta era buscar aferir a taxa de sobrevivência das *startups* a partir da existência de instituições de ensino e de instituições de apoio. Mudamos a abordagem após nos depararmos com essa limitação temporal.Outra limitação identificada foi haver bibliografia escassa relacionada a *startups* e aceleradoras, não só no Brasil. Por ser uma temática recente, ainda há poucos estudos acadêmicos relacionados ao assunto, principalmente quando aliamos o tema startups com técnicas de análise de dados com uso de estatística espacial. Consideramos isso como uma limitação, mas também como uma oportunidade de aprofundarmos no tema e contribuir para mudar esse cenário.

Ainda, identificamos que o modelo que desenvolvemos foi feito com base em variáveis com alta correlação entre si. Como o fato aconteceu, nosso modelo não foi capaz necessariamente de discernir a relação de cada uma das variáveis que estávamos tentando explicar. Como já abordamos, o componente geográfico, também presente nas variáveis econômicas, acabou por contribuir para que a utilização do método SAR não melhorasse os índices do modelo.

Identificamos, de forma não exaustiva, oportunidades para o desenvolvimento de trabalhos futuros. A metodologia CRISP-DM prega como primeira e uma das mais importantes fases a etapa de entendimento do negócio. Recomendamos que em trabalhos futuros realize-se pesquisas de campo com as *startups* e entidades de apoio visando a obtenção de novas variáveis que não foram contempladas em nosso estudo. Compreender os perfis de sócios, investidores e equipes de trabalho é fundamental para avaliar a criação e manutenção de startups.

Além de pesquisas de campo, utilizar outras variáveis disponíveis bem como outras técnicas de análise, como, por exemplo, Árvore de Decisão para identificar distâncias entre as *startups* e instituições de ensino e outras instituições de apoio e Teoria de Grafos para identificar relacionamentos. Após isso, seria interessante realizar um estudo comparando diversas técnicas de exploração de dados para análise do fenômeno.

Observamos na base de dados de *startups* utilizada nesse trabalho uma variedade de nichos de negócio: Desde *startups* da área de cooperativas agroindustriais até *startups* de tecnologia da informação ou que estão fortemente baseadas na tecnologia da informação. Consideramos ser estudos pertinentes aqueles que busquem avaliar quais tipos de *startups* possuem maior taxa de sobrevivência, bem como se há concentração de *startups* no Brasil por nichos ou por vocação das diversas regiões.

Identificamos a pertinência de avançar com estudos sobre o tema para auxiliar potenciais investidores na prospecção de oportunidades de negócio. Como o tema startups abrange diversos setores da economia como, por exemplo, agronomia, serviços bancários, mobilidade urbana, etc. seria importante a realização de uma análise segmentada, seja por região de interesse, seja por ramos de atividade. Um possível caminho pode ser trabalhar com dados de entidades setoriais como a Confederação Nacional da Indústria (CNI), Confederação Nacional do Comércio (CNC) e Serviço Nacional de Aprendizagem Industrial (Senai) ou mesmo agregadores nacionais de informação como o Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC).

Por fim, apontamos como relevante a realização de estudos de análise de estatística espacial que subsidiem a tomada de decisão através da predição de locais ótimos para a abertura de novas entidades de apoio ou a identificação de áreas de influência com alto potencial para o desenvolvimento de *startups*, o que pode auxiliar órgãos governamentais no implemento de

políticas públicas que incentivem a criação de incubadoras e aceleradoras para potencializar o desenvolvimento de regiões. Tais achados permitem endereçar melhor gestão e incentivos para a sustentação e aumento da oferta de *startups* em todo o território nacional.

REFERÊNCIAS BIBLIOGRÁFICAS

Aguilhar, L. (2014, Março) Aceleradoras de startups ganham força no Brasil, mas precisam provar eficiência. *Estado de São Paulo*. Recuperado de https://link.estadao.com.br/noticias/geral,aceleradoras-de-startups-ganham-forca-no-brasil-mas-precisam-provar-eficiencia,10000031831.

Anprotec – Associação Nacional de Entidades Promotoras de Empreendimentos Inovadores. (2016) Estudo de impacto econômico: segmento de incubadoras de empresas do Brasil. Brasília: ANPROTEC/SEBRAE.

Dullius, A. C. & Schaeffer, P. R. (2016). A capacidade de Inovação em *Startups*: contribuições para uma trajetória de crescimento. *Revista Alcance (Online), v. 23*, 34-50.

Instituto Brasileiro de Geografia e Estatística – IBGE. (2017). *Divisão Regional do Brasil em Regiões Geográficas Imediatas e Regiões Geográficas Intermediárias: 2017*. Rio de Janeiro:

Francisco, E. R., Aranha, F., Zambaldi, F., Goldszmidt, R. Electricity Consumption as a Predictor of Household Income: a Spatial Statistics Approach. Advances in Geoinformatics. pg. 267-282. GEOINFO 2006. Campos de Jordão, Brazil, 2006.

IBGE. Recuperado de https://biblioteca.ibge.gov.br/visualizacao/livros/liv100600.pdf

Minatogawa, V. L. F. (2013). *Estudo e Adaptação de um Método de Avaliação do Desempenho de Modelos de Negócios em uma Organização Startup***.** Recuperado de http://repositorio.unicamp.br/jspui/bitstream/REPOSIP/265629/1/Minatogawa_ViniciusLuizFerraz_M.pdf

Nogueira, V. S. & Oliveira, C. A. A. (nov./fev. 2014/2015). Causa da mortalidade das *startups* brasileiras: como aumentar as chances de sobrevivência no mercado. Nova Lima, *DOM: v.9, n. 25*, 26-33.

Ribeiro, A. T. V. B., Plonski, G. A. & Ortega, L. M. (2015). *Um fim, dois meios: aceleradoras e incubadoras no Brasil. In Inovação para além da tecnologia*. São Paulo: ALTEC.

Ries, E. (2012). *A Startup Enxuta: Como os empreendedores atuais utilizam inovação contínua para criar empresas extremamente bem-sucedidas*. São Paulo: Leya Editora.

Serviço Brasileiro de Apoio às Micro e Pequenas Empresas Nacional - SEBRAE-NA (2017). *Relatório Especial: O Financiamento das MPE no Brasil*. Recuperado https://bibliotecas.sebrae.com.br/chronus/ARQUIVOS_CHRONUS/bds/bds.nsf/843a3cb8bd1901b9dec713eb3cdbb260/$File/7754.pdf

Silva, E. E. da. (2017). *Ciclo de vida das empresas startups*. Campo Limpo Paulista, SP: FACCAMP.

Torres, J. (2017). *Guia da Startup: Como startups e empresas estabelecidas podem criar produtos de software rentáveis.* 2. ed. São Paulo: Casa do Código.

Vergara, S. C. (2005). *Métodos de pesquisa em administração*. 1ª edição. São Paulo: Editora Atlas.

Weiblen, T. & Chesbrough, H.W. (2015). Engaging with *Startups* to Enhance Corporate Innovation. *California Management Review, vol. 57(2)*, 66-90.