Plan Overview

A Data Management Plan created using DMPTool

Title: Minha SP

Creator: Arthur Gusmao

Affiliation: Non Partner Institution

Principal Investigator: Arthur Gusmao

Data Manager: Arthur Gusmao

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: 30476

Template: NSF-GEN: Generic

Last modified: 12-18-2017

Minha SP

Duas diferentes fontes de dados serão utilizadas durante o projeto: IBGE e PRODAM.

IBGE

A primeira fonte de dados utilizada no desenvolvimento desta pesquisa é o IBGE (Institudo Brasileiro de Geografia e Estatística). Os dados coletados desta fonte foram gerados no último Censo Demográfico realizado no Brasil, no ano de 2010.

O Censo Demográfico é a mais complexa operação estatística realizada por um país, quando são investigadas as Figura 1. Todos os setores censitários do estado de São Paulo, plotados sobre o mapa do Brasil. A imagem foi produzida utilizando o software QGIS Las Palmas v2.18 caraterísticas de toda a população e dos domicílios do território nacional.

Os Censos Demográficos, por pesquisarem todos os domicílios do país, constituem uma valiosa fonte de referência para o conhecimento das condições de vida da população. Os dados deste arquivo, que estão agregados por setor censitário, compreendem características dos domicílios particulares e das pessoas que foram investigadas para a totalidade da população.

Um setor censitário é a menor divisão territorial adotada pelo IBGE. Juntos, todos os setores censitários formam o território brasileiro. Nesta pesquisa, foram utilizados apenas os setores censitários que se combinam para formar o estado de São Paulo.

Além das variáveis de identificação geográfica, as informações em nível de setor estão distribuídas em planilhas, com cerca de 3.000 variáveis, que abrangem as seguintes características da população residente: sexo, idade, cor ou raça, condição no domicílio; pessoas responsáveis pelo domicílio; alfabetização; registro de nascimento das crianças de até 10 anos de idade; e características dos domicílios particulares.

Soma-se aos conjuntos de dados um shapefile (um arquivo vetorial) que possibilita a criação de um layer de visualização sobre mapas e também cálculos georeferenciados utilizando ferramentas de Geographic Information System (GIS). Tanto os conjuntos de dados quanto os shapefiles estão disponíveis para download no website do IBGE (hhtp://www.ibge.gov.br/home/).

PRODAM

A segunda fonte de dados utilizados é a PRODAM (Empresa de Tecnologia da Informação e Comunicação do Município de São Paulo). Embora a PRODAM disponibilize dados apenas para o município de São Paulo, seus dados foram incorporados no estudo.

Dois tipos de informação foram coletadas do portal da PRODAM. A primeira é referente às favelas do município de São Paulo. Este conjunto de dados contém os perímetros das favelas cadastrados, contendo nome, endereço, ano de implantação e estimativa do número de domicílios, entre outros. Excluem-se, aqui, os dados referentes aos Núcleos Urbanizados da Cidade de São Paulo, que contém favelas que já possuem infraestrutura de água, esgoto, iluminação pública, drenagem e coleta de lixo.

E o outro tipo de informação é referente aos moradores de rua da cidade de São Paulo, e traz pontos de concentração de população em situação de rua com informação de sexo, idade e data de abordagem.

Assim como no caso do IBGE, os dados da PRODAM estão disponíveis para download no próprio site (http://dados.prefeitura.sp.gov.br/pt_PT/), sendo necessário obter os datasets e os shapefiles.

DADOS GERADOS

Ao final do projeto, a partir dos dados utilizados inicialmente, serão produzidos um número (ainda indeterminado) de clusters para cada região (polígonos) do estado de São Paulo. Estes dados serão disponibilizados para toda a comunidade, e será gerado um DOI para cada conjunto. Ambas etapas serão feitas através do repositório online Figshare (http://figshare.com).

Todos os dados referentes a regiões (polígonos) do estado de São Paulo deverão ser disponibilizados na forma de shapefiles: arquivos vetoriais que permitem criação de camadas de visualização. O intuito é que dessa forma ferramentas de Geographic Information System (GIS) possam ser usadas de modo uniforme em todos os conjuntos utilizados.

Como os shapefiles são constituídos de diferentes arquivos, é necessário que todo e qualquer conjunto disponibilizado como shapefile apresente todas as extensões de arquivos:

- .dbf
- .prj
- .qpj
- .shp
- .shx

Além disso, por se tratarem de arquivos relativamente grandes, os shapefiles poderão ser diponibilizados em forma compactada. Para isso, o formato zip deverá ser utilizado.

Os arquivos que não correspondem a regiões (polígonos) e podem ser representados na forma de tabelas deverão ser disponibilizados no formato csv padrão (arquivos separados por vírgulas).

Em casos em que os arquivos csv forem demasiadamente grandes, poderão, assim como os shapefiles, serem disponibilizados em formato compactado (zip). Entretanto, este tipo de formato deverá ser evitado porque ele dificulta a exibição dos dados de forma direta em ferramentas como o Figshare, prejudicando o acesso ao conteúdo.

Todos os dados gerados durante o projeto serão disponibilizados livremente na internet, e poderão ser acessados, por qualquer pessoa ou comunidade, através do repositório referente ao projeto, que ficará na ferramenta Figshare.

Os scripts utilizados para processamento dos dados, por sua vez, serão armazenados na ferramenta online GitHub, onde um repositório será criado especificamente para o projeto. Lá, qualquer pessoa, sendo usuário ou não do GitHub, conseguirá livre acesso aos scripts.

Tanto os dados quanto os scripts serão disponibilizados junto com a publicação do artigo.

O Figshare gera automaticamente um DOI (Digital Object Identifier), um identificador persistente utilizado para identificar unicamente os objetos. Dessa forma, qualquer pessoa ou comunidade científica poderá reutilizar ou redistribuir livremente os dados, desde que citados da devida forma, conforme delineado no repositório.

A licensa sob a qual os scripts serão disponibilizados será uma licensa MIT. Ela é uma licensa gratuita, que se originou do instituto norte americano Massachusetts Institute of Technology. É permissiva, com poucas restrições, apresentando portanto uma excelente compatibilidade. Um template para a licensa pode ser encontrado em

(https://opensource.org/licenses/MIT).

Os dados ficarão disponíveis no repositório Figshare indefinidamente. Entretanto, os autores se responsabilizam por manter os dados acessíveis no período de um ano, tomando as devidas medidas para garantir o acesso dos dados. No caso em que o Figshare não esteja mais disponível ou o acesso torne-se pago, os autores têm a liberdade de poder escolher alguma outra forma para armazenamento e disponibilização.

Com relação aos scripts, garantiremos que os mesmos ficarão acessíveis por um período mínimo de um ano no devido respositório, no GitHub, com a exceção do caso em que o GitHub encerrar suas atividades ou começar a exigir pagamento para os serviços prestados. Nesta situação, os autores ficam livres para escolher um novo repositório para armazenamento.
