ABORDAGEM DE APLICAÇÃO DE ENGENHARIA DE SISTEMAS DE SISTEMAS A GRANDE VOLUME DE DADOS

Ernesto Rademaker Martins
Centro de Análises de Sistemas Navais (CASNAV)
Praça Barão de Ladário, s/nº Ed. 23 do AMRJ – Centro – Rio de Janeiro – RJ
Engenharia de Defesa - Instituto Militar de Engenharia
rademaker@marinha.mil.br

Marcio Roberto Galhano
Instituto Brasileiro de Geografia e Estatística (IBGE)
Avenida Beira Mar, 436 – Glória – Rio de Janeiro – RJ
mrgalhan@gmail.com

Marcos dos Santos
Centro de Análises de Sistemas Navais (CASNAV)
Praça Barão de Ladário, s/nº Ed. 23 do AMRJ – Centro – Rio de Janeiro – RJ
marcossantos@marinha.mil.br

Cícero Roberto Garcez
Engenharia de Defesa - Instituto Militar de Engenharia (IME)
Praça General Tibúrcio, 80 – Urca – Rio de Janeiro – RJ
garcez@ime.eb.br

RESUMO

Com objetivo de criar uma metodologia de modelagem em grafos para representação de algumas relações com base em dados de pesquisa do CD2010 através da ferramenta denominada SIDRA (Sistema do IBGE para Recuperação Automática), que disponibiliza para o público dados desagregados, em formato de tabelas demonstrativas da realidade brasileira. Portanto, será apresentada uma metodologia baseada na Engenharia de Sistemas de Sistemas para representação de uma rede de nós e arestas, usando como fonte de dados a relação de características físicas e sociais (sexo, cor, raça, situação domiciliar e grupos de idade) por localização geográfica (entre bairros e unidades federativas).

Palavra-chave: Metodologia; Grafos; Engenharia de Sistemas de Sistemas; SIDRA.

ABSTRACT

In order to create a graphs modeling methodology to represent some relationships based on research data of 2010 Brazilian Demographics Census (CD2010) by the tool called SIDRA (IBGE Automatic Retrieval System) that provides to the public a disaggregated data in tables demonstrating Brazilian reality. Therefore, the objective will be to present a methodology based on Systems of Systems Engineering to represent a network of nodes and edges, using as data source the relationship of physical and social characteristics (gender, color, race, home situation and age groups) by geographical location (between neighborhoods and federative units).

Keywords: *Methodology*; Graphs; *Systems of Systems Engineering*; SIDRA.

Como Citar:
RADEMAKER, E. M.; GALHANO, M.; SANTOS, M.; GARCEZ, C. R.; Metodologia de Aplicação de Engenharia de Sistemas de Sistemas à Grande Volume de Dados. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

Este trabalho se justifica ao propor apresentar uma metodologia na aplicação da Engenharia de Sistemas de Sistemas no intuito de gerar uma visão espacial ou especial dinâmica, baseada em grafos, de dados que normalmente são disponibilizados apenas em tabelas. A visão em grafos facilita o entendimento de relacionamentos e, ainda, evidencia outras relações não naturais existentes. Esta não evidente relação pode ser a aurora de novos estudos prospectivos relacionais entre dados que aparentemente poderiam não apresentar relações de quaisquer espécies.

A associação da conexidade da Teoria dos Grafos com dados de grande volume (*Big Data*) foi recentemente usada pela empresa *Cambridge Analytics*, para supostamente ajudar na eleição do então candidato Donald Trump nas eleições presidenciais americanas (GUZDIAL, M.; LANDAU, S. 2018). Essas questões intrigantes de relacionamentos não evidentes e o sucesso na exploração de dados não revelados se justificam como relevante pela sua continuidade no âmbito da pesquisa científica.

2. CONTEXTO

A apresentação de resultados de pesquisa a partir dos dados da CD2010 sobre as características da população em foco e de seus domicílios foram disponibilizados em grafo. As notações foram alteradas para facilitar a compreensão por parte de pesquisadores do IBGE. Para tratar a conexidade, foram usadas as definições dadas por Boavetura (2006, p.31) onde define conexidade como à possiblidade de passagem de um vértice a outro através das ligações existentes. Ainda este autor diz que a ideia de passagem se refere a “atingibilidade” (BOAVENTURA, 2006, p.7). Para este trabalho o conceito de “atingibilidade” nada mais é do que a relação entre vértices.

As arestas conectam indivíduos com características comuns, ou seja, as arestas serão as características comuns que ligam os indivíduos.

Assim, logicamente, os vértices representam os indivíduos, com suas características

associadas por arestas. Cada aresta provê uma característica comum entre indivíduos. Essa característica pode ser região geográfica, local de nascimento, idade, sexo, grau de escolaridade, cor da pele, entre outras (BRANDT, 2018).

3. PROBLEMÁTICA, HIPÓTESE E OBJETIVOS

Os dados do IBGE, mais precisamente do CD2010, estão relacionados diretamente em uma grande planilha eletrônica, com associações tabulares, apresentando as características da população brasileira. Contudo os dados são numerosos, na ordem de 648 linhas versus 09 colunas para cada linha, o que dificulta uma avaliação tanto da correção dos dados, quanto da relação entre eles. Então, caso se faça uma relação gráfica, baseada em conceitos de grafos, acredita-se que os relacionamentos possam ser melhor compreendidos, podendo indicar uma correlação de alguma ordem entre eles (COWLS, J; SCHROEDER, R., 2015).

A correlação não necessariamente nos dá uma condição de causalidade, mas nos dá uma indicação de onde os estudos devem começar e porque há essa relação. Sugere-se que comece com uma correlação linear de Pearson, verificando se ela existe e depois seguir para outras correlações não lineares, que não serão foco deste trabalho (NÆSS, P; PETERS, S; STEFANSDOTTIR, H; STRAND, A, 2018).

O objetivo deste trabalho se resume a apresentar a metodologia sistêmica de criação de um grafo com relações existentes em uma tabela do Censo Demográfico de 2010, realizado no Brasil, com uma quantidade limitada de correlações diretas, a partir de matriz de valores coletados no CD 2010.

4. JUSTIFICATIVA

Justifica-se o trabalho pela necessidade de criar uma metodologia científica, baseada na Engenharia de Sistemas para análise de grande base de dados estruturada matricialmente. Trabalhos semelhantes foram conduzidos em outros países, com outras bases de dados, mas nenhum trabalho foi desenvolvido especificamente para o Censo Brasileiro de 2010, utilizando uma ferramenta de visualização em grafos, como o *software* gratuito *Gephi* (AMAT, 2014).

No que pese as relações serem dinâmicas e complexas, este trabalho se limitou a estudar uma matriz específica, estática no tempo, tendo tentado posteriormente fazer comparações.

5. CONTRIBUIÇÃO

A Engenharia de Sistemas tem presente em sua estrutura a capacidade de trabalhar com grande volume de dados, quando sistematiza o processo de coleta, validação e análise de dados. O cruzamento de técnicas como a Engenharia de Sistemas, transdisciplinar (GARCEZ, 2019), com a Pesquisa Operacional, com a ferramenta Teoria dos Grafos, intradisciplinar (WHITE, D; DONALDSON, W; LAWRIE, N., 1969), traz a possiblidade de estudos profundos e relevantes para o desenvolvimento de novas fronteiras de pesquisa e análise, no âmbito da ciência de dados, em especial, quando se trata de Big Data.

6. ENGENHARIA DE SISTEMAS DE SISTEMAS

Existe uma complexidade intrínseca na análise de grande volume de dados, pois esta análise necessita que haja processos hierarquizados no tempo. Com as tarefas organizadas em ordem sequencial, inicia-se a utilização das ferramentas, tanto da Engenharia de Sistemas, como da Pesquisa Operacional, em seu tempo correto, para que os resultados sejam apresentados no tempo e com a qualidade esperada/adequada.

Assim, este trabalho sugere adoção de técnicas sistêmicas de modo sequencial, aplicado ao sistema e aos sistemas desse sistema:

6.1 Diagramação do ciclo de vida;
6.2 Diagrama de contexto;
6.3 EAP – Estrutura Analítica de Projetos;
6.4 Diagrama de Domínio – de atividades;
6.5 PERT – *Program Evaluation and Review Technique*;
6.6 Análise de massa de dados – SIDRA; e
6.7 Criação de grafo com dados provenientes da SIDRA.

6.1. CICLO DE VIDA

O ciclo de vida consiste, neste estudo em todo processo construtivo, desde sua concepção até seu descarte, com o armazenamento de dados, para estudos futuros.

O Ciclo de vida desenvolvido para este trabalho foi baseado no gerenciamento do ciclo de vida de equipamento a bordo de navios (MARTIN, B. *et al*, 2018), pelo fato de navios serem sistemas complexos. Um exemplo disso é o diagrama da Figura 1.

Como parte complementar do curso, foi feita uma visita técnica à Fragata Rademaker (F-49), onde os alunos do curso de Engenharia Defesa, do Instituto Militar de Engenharia, puderam conhecer a complexidade do sistema de sistemas de um navio de guerra da Marinha do Brasil, apresentado pelo Comandante, o senhor Capitão de Fragata Leonardo Mesquita Araújo.

Esta visita foi a confirmação prática da utilidade dos ensinamentos de sala de aula. Nesta ocasião foi apontada a fase operacional como parte integrante do ciclo de vida de um navio de guerra.

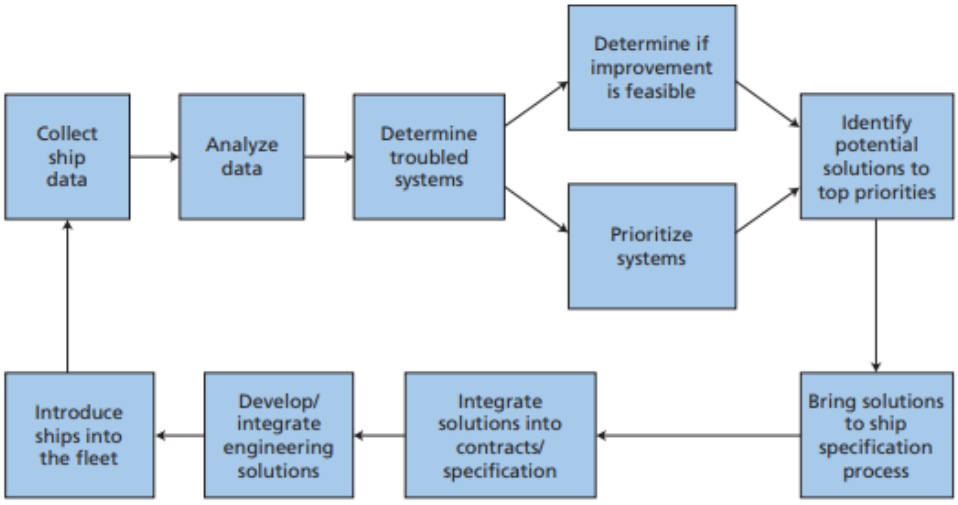

Figura 1 - "*Construction-Modernization Feedback Loop*". Fonte: MARTIN, B. *et al*, 2018, p.03

O ciclo de vida do projeto de Engenharia de Sistemas de Sistemas de grande volume de dados do CD 2010 está apresentado na Figura 2.

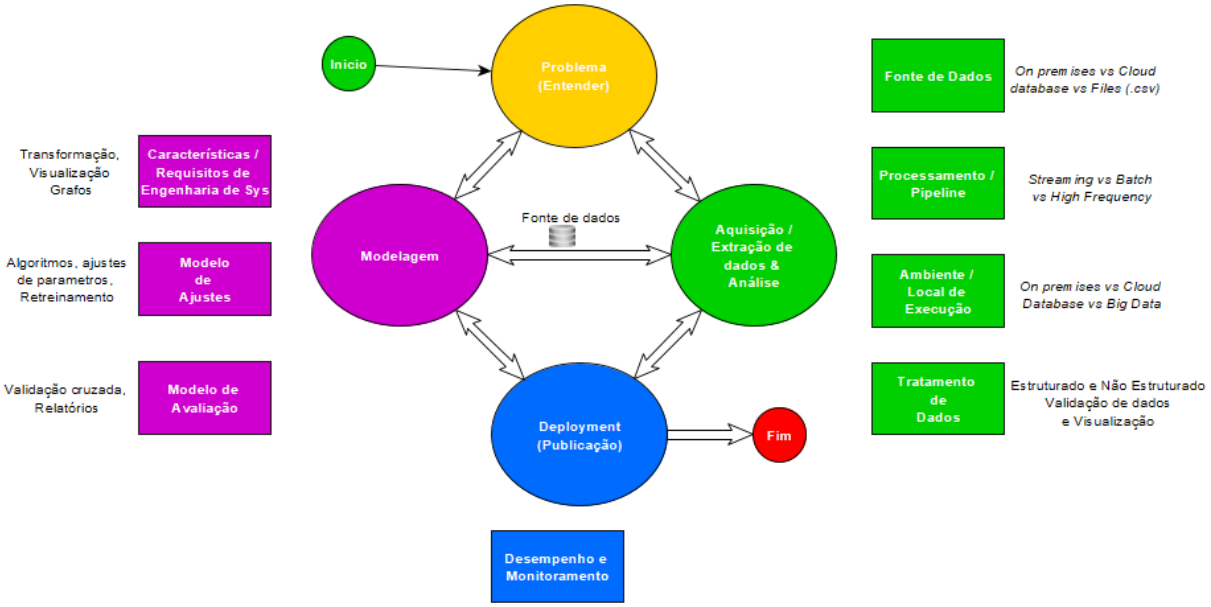

Figura 2 – Ciclo de vida do projeto de Engenharia de Sistemas de Sistemas de grande volume de dados do CD 2010. Fonte: autores (2019).

6.2. DIAGRAMA DE CONTEXTO

O Diagrama de contexto é um modelo de ferramenta estrutural para entender as conexões existentes entre sistemas, por meio de um diagrama esquemático.

No caso específico deste trabalho, buscou-se apresentar as ligações e interações entre os vários sistemas, assim como a colaboração mútua entre elas para o encontro de relações não convencionais entre vértices e arestas dos grafos, ou seja, relações não triviais com as informações disponíveis. O diagrama de contexto apresenta claramente uma construção *middle out*, representado pelos sistemas no interior do círculo maior, fornecendo como resultado a análise relacional a partir da fonte de dados do IBGE, o Censo Demográfico de 2010. Outros sistemas necessários para que o sistema *middle out* funcionasse, foi a necessidade de adoção de dois outros sistemas externos: requisitos, com as variáveis populacionais e a classificação de dados, com a organização tabular, conforme apresentado na Figura 3.

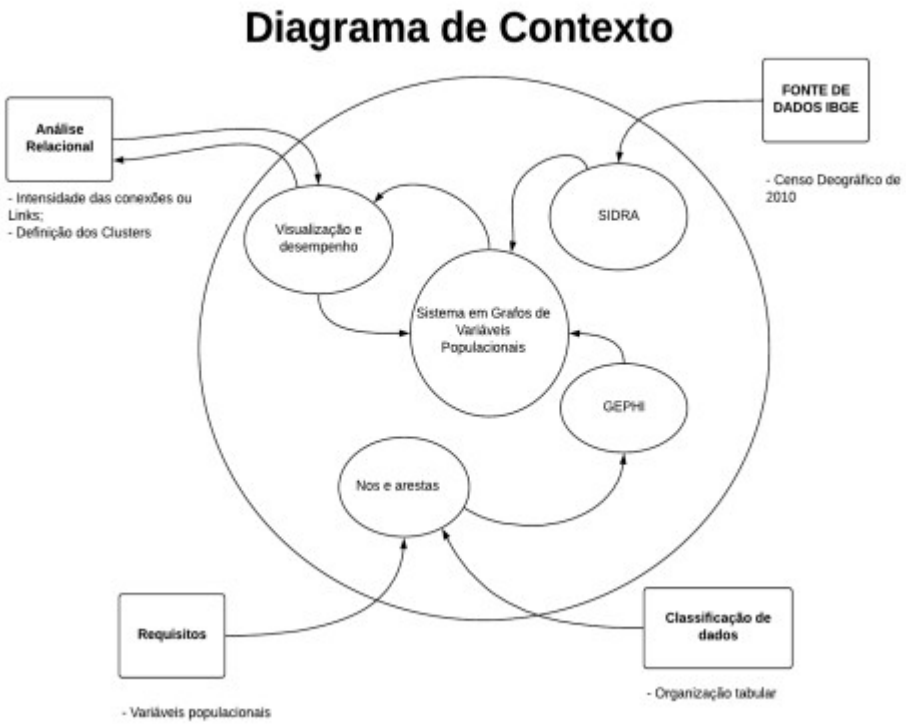

Figura 3 – Diagrama de contexto do Engenharia de Sistemas de Sistemas aplicada à visão espacial em Grafos das Características da População Brasileira. Fonte: autores (2019).

6.3. EAP – ESTRUTURA ANALÍTICA DE PROJETOS

A estrutura analítica do projeto é a representação gráfica da gestão do processo de construção de um sistema, que deriva de outro sistema: Sistema de Recuperação Automática de Dados – IBGE (SIDRA). Esta estrutura analítica demonstra por meio de diagrama de blocos as etapas de construção do modelo, adicionando, de forma perspicaz, a curva ABC, o que proporciona a possiblidade de relacionar a produtividade versus recursos, conforme apresentado na Figura 4.

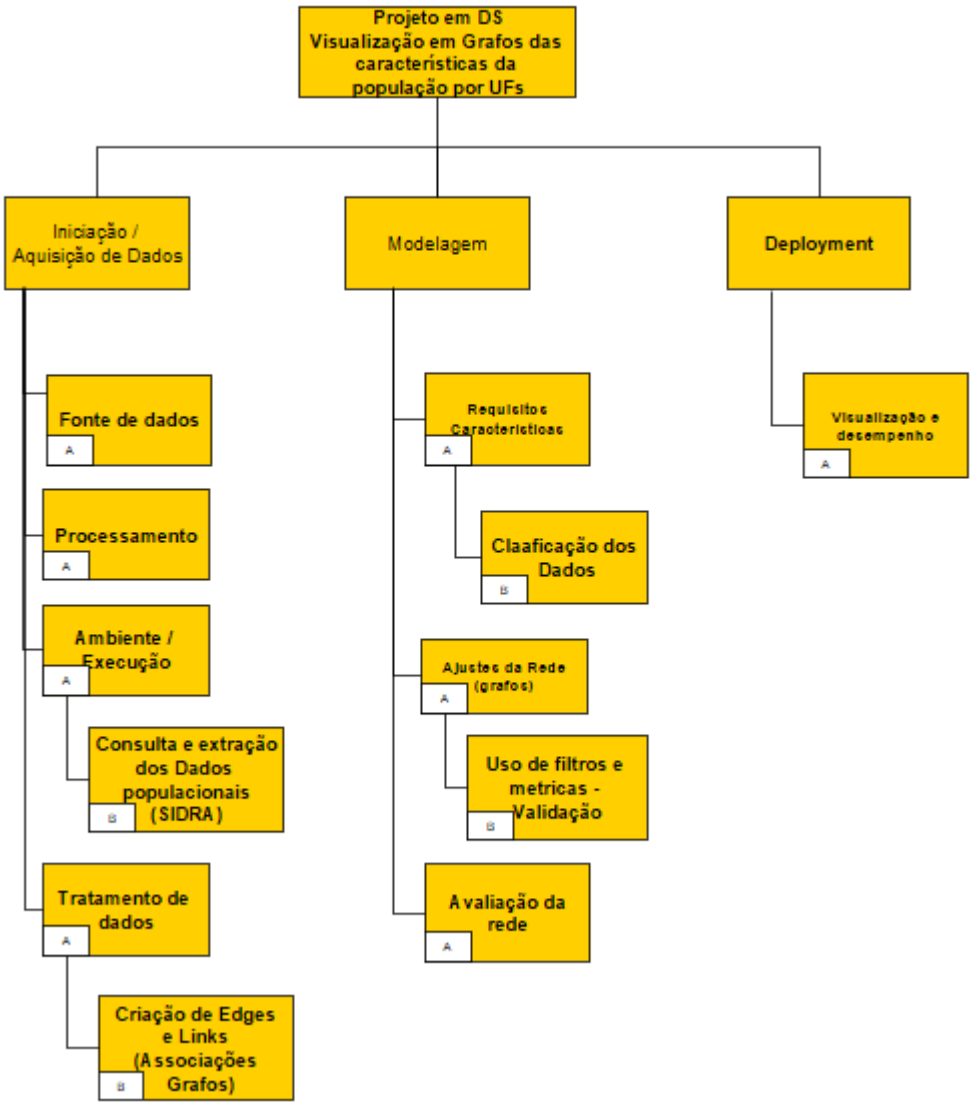

Figura 4 – Estrutura Analítica do Projeto de Engenharia de Sistemas de Sistemas aplicada à visão espacial em Grafos das Características da População Brasileira. Fonte: autores (2019).

6.4. DIAGRAMA DE DOMÍNIO DE ATIVIDADES

Este diagrama nada mais é do que o gráfico de *Gantt*. Ele permite organizar as tarefas e recursos no tempo, permitindo a identificação do caminho crítico do projeto, a partir de um conjunto de tarefas cujos tempos de execução, se não cumpridos, podem comprometer o projeto. Este diagrama auxilia no planejamento e execução das tarefas de um sistema complexo, com outros sistemas relacionados. Este Diagrama está apresentado na Figura 5.

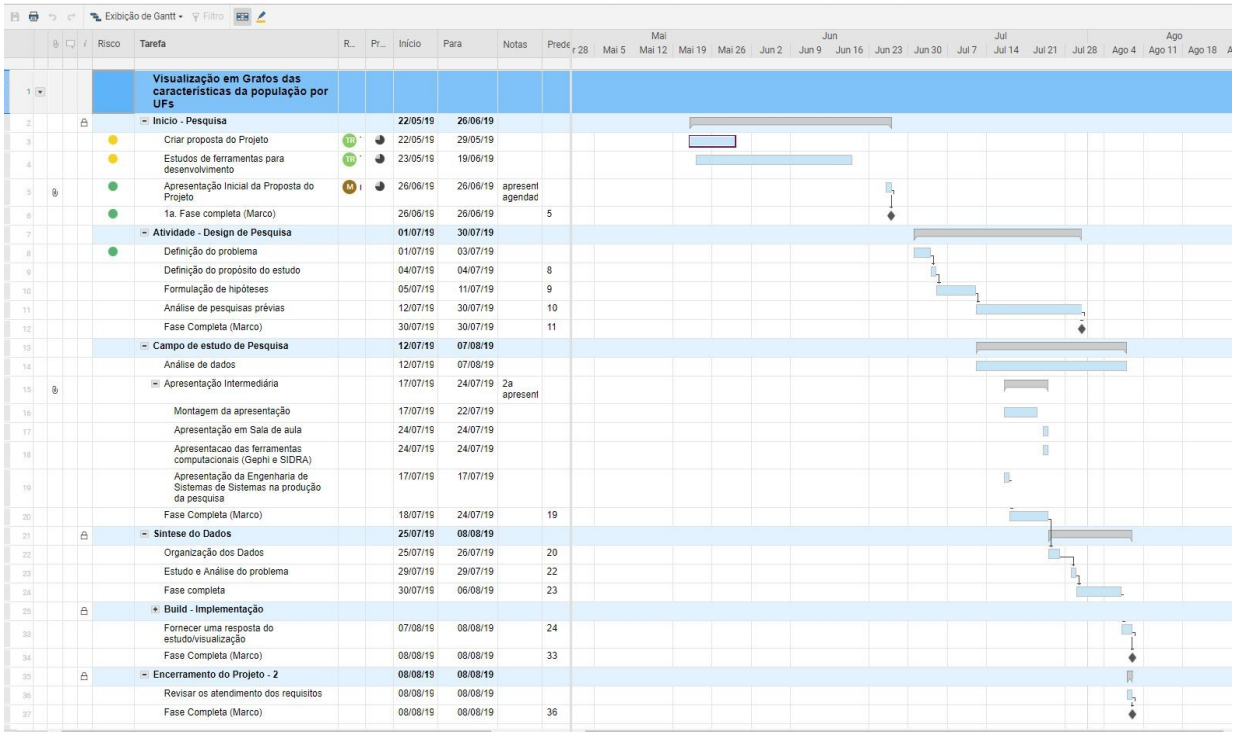

Figura 5 – Diagrama de atividades do Projeto de Engenharia de Sistemas de Sistemas aplicada à visão espacial em Grafos das Características da População Brasileira. Fonte: autores (2019).

6.5. DIAGRAMA PERT – *PROGRAM EVALUATION AND REVIEW TECHNIQUE*

O diagrama PERT permite encontrar o caminho crítico para as atividades de um projeto. Este diagrama identificou que duas atividades paralelas em uma fase do projeto faziam parte do caminho crítico. Depois de uma análise mais apurada, percebeu-se que o caminho crítico pertencia a apenas uma das atividades, ficando o modelo devidamente registrado como um único caminho crítico, sem bifurcação intermediária. O diagrama PERT deste projeto de modelagem de metodologia está apresentado na Figura 6.

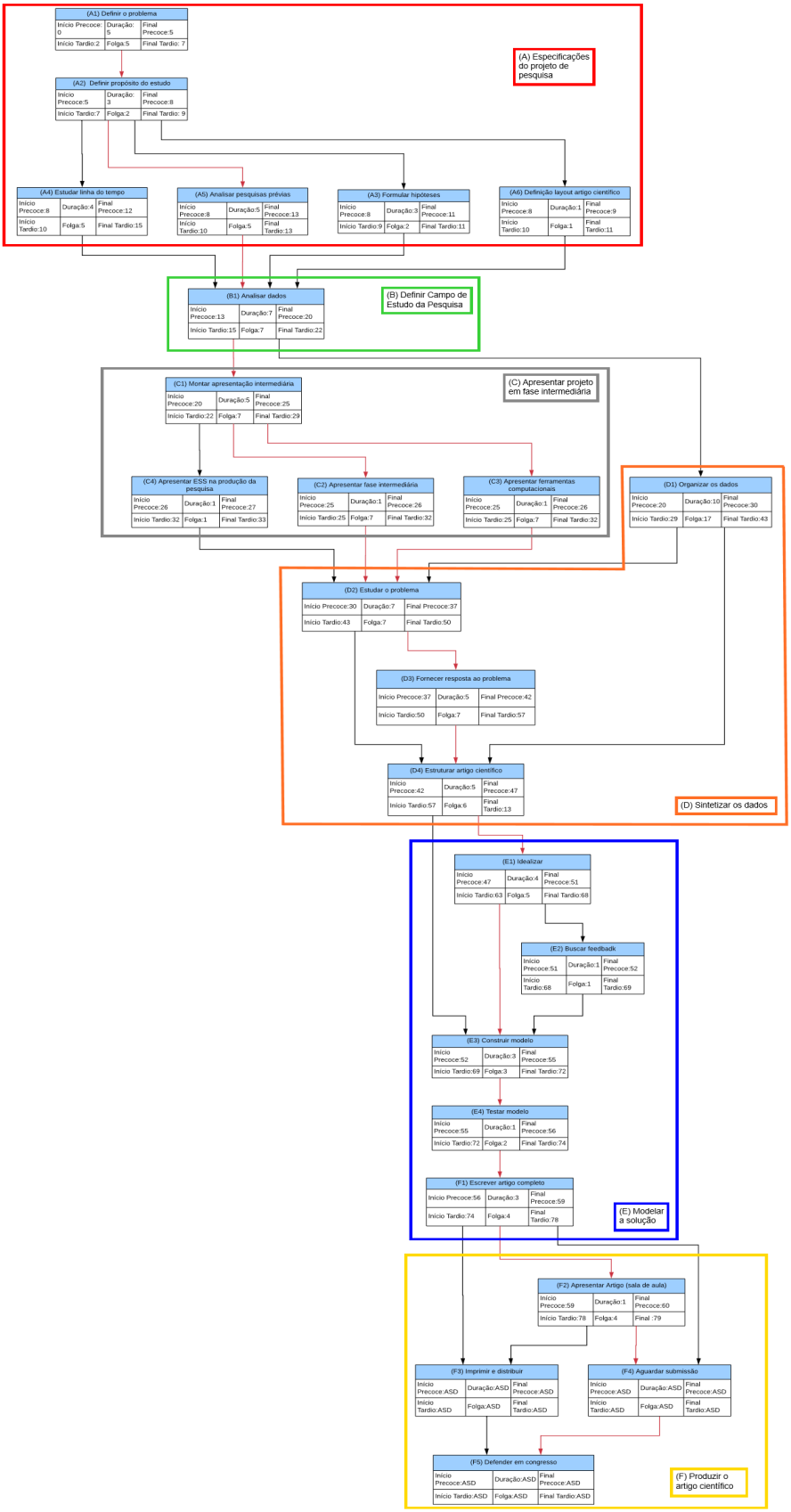

Figura 6 – PERT do Projeto de Engenharia de Sistemas de Sistemas aplicada à visão espacial em Grafos das Características da População Brasileira. Fonte: autores (2019).

Diferente do comumente apresentado, o diagrama possui grandes blocos, que indicam fases do processo, ou seja, os blocos devem ser totalmente concluídos antes do início de atividades do bloco subsequente. O caminho crítico está indicado com a linha vermelha, na Figura 6.

6.6. ANÁLISE DE MASSA DE DADOS

Os dados analisados são provenientes da uma das tabelas do SIDRA que informa o quantitativo de domicílios particulares permanentes por situação do domicílio, e existência de compartilhamento da responsabilidade pelo domicílio, por grupos de idade e sexo e unidades da federação.

Esta análise prévia dos dados e organizações em tabela permitiu a inserção deles, de forma tabular, devidamente catalogados no software *Gephi*, para análise gráfica.

6.7. GRAFO RESULTANTE DA MASSA DE DADOS - SIDRA

A visualização, parte final do processo, foi provida com a conversão da massa de dados em uma matriz, que por sua vez, foi inserida no software *Gephi*, para visualização dos relacionamentos, triviais ou não. Os relacionamentos podem ser vistos na Figura 7.

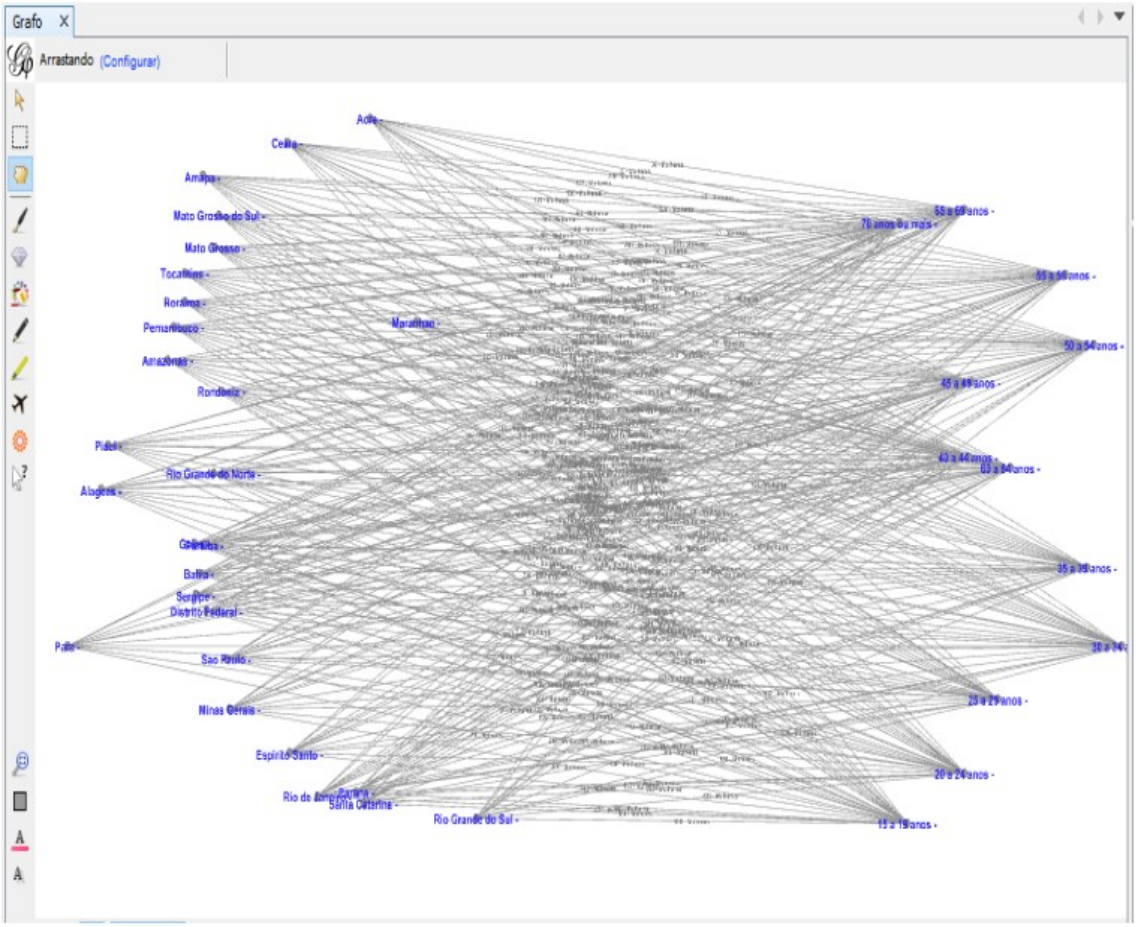

Figura 7 – Grafo provido pela massa de dados SIDRA. Fonte: autores (2019).

Obviamente que este artigo não tem a pretensão de liquidar o assunto, muito pelo contrário, ele é provocador de construção analítica a partir de visualização gráfica, a partir da metodologia desenvolvida.

7. CONCLUSÃO

A Engenharia de Sistemas de Sistemas apresenta grande valor quando se trata da estruturação de modelo para solução de problemas complexos, principalmente no que trata de grande volume de dados. A estruturação promovida pelas técnicas aplicadas devidamente no tempo, com o auxílio das ferramentas de Pesquisa Operacional, permitiu uma modelagem adequada do processo, culminando neste trabalho apresentado como metodologia preparatória para produção de trabalho científico de Data Science.

Curiosamente, no decorrer do trabalho a modalidade *middle out* se sobressaiu e, por fim, dominou o contexto situacional, evidenciado no Item 6.2 Diagrama de Contexto, o que, em princípio, não era esperado.

A sequência de tarefas foi construída naturalmente, conforme o trabalho evoluía. Contudo, não há como afirmar que seja a única sequência adequada, ou até mesmo, a mais adequada. Pode-se afirmar somente que este processo teve êxito ao chegar no tempo esperado, à modelagem de sistema de sistemas complexos, voltado para grande volume de dados.

8. SUGESTÃO DE TRABALHOS FUTUROS

Este trabalho se limitou a modelar o processo de tarefas, baseado na Engenharia de Sistemas de Sistemas, em processo de construção da metodologia de pesquisa para produção de trabalho científico de grande volume de dados.

Sugere-se, em trabalhos futuros, desenvolver metodologia para a Engenharia de Sistemas de Sistemas no trato de cenários dinâmicos, ainda mais complexos, com a utilização da ferramenta DyCoNet (KAUFFMAN, J. et al, 2014).

9. REFERÊNCIAS BIBLIOGRÁFICAS

AMAT, C. B. **Análisis y visualización de redes con Gephi.** REDES: Revista Hispana para el Análisis de Redes Sociales, Vol.25(1), pp.201-209, Junho 2014.

BOAVENTURA NETTO, P. O. **Grafos: teoria, modelos e algoritmos**. São Paulo. Editora Edgard Blücher, 4ª Ed. 2006.

BRANDT, M.B. **Modelo de dados abertos conectados para informação legislativa Mariana.** Informação & Sociedade, Vol.28(2). 2018.

COWLS, J ; SCHROEDER, R. **Causation, Correlation, and Big Data in Social Science Research.** Policy & Internet, Vol.7(4), pp.447-472, Dezembro 2015.

GARCEZ, C.R. **Notas de Sala de Aula - Engenharia de Sistemas de Sistemas**. Instituto Militar de Engenharia, Rio de Janeiro, 2019.

GUZDIAL, M; LANDAU, S. **Programming Languages, and Analyzing Facebook's Failure.** Association for Computing Machinery. Communications of the ACM, Vol.61(6), p.8, Jun 2018.

KAUFFMAN, J; KITTAS, A; BENNETT, L; TSOKA,S. **DyCoNet: a Gephi plugin for community detection in dynamic complex networks.** PLoS ONE, Vol.9(7), p.e101357. 2014.

MARTIN, B; YARDLEY, R; PARDUE, P; TANNEHILL, B; WESTERMAN, E; DUKE, J. **An Approach to Life-Cycle Management of Shipboard Equipment**. Santa Monica, CA: RAND Corporation, 2018. Disponível em: https://www.rand.org/pubs/research_reports/RR2510.html. Acesso em: 26AGO2019.

NÆSS, P; PETERS, S; STEFANSDOTTIR, H; STRAND, A. Causality**, not just correlation: Residential location, transport rationales and travel behavior across metropolitan contexts.** Journal of Transport Geography, Vol.69, pp.181-195, Maio 2018.

WHITE, D; DONALDSON, W; LAWRIE, N. **Operational Research Techniques**. Business Books Limited. Londres. 1969.