



## 3º Encontro Internacional de Dados, Tecnologia e Informação

Perspectivas e Interdisciplinaridades  
em Ciência da Informação

Marília | SP

04 a 06 de outubro de 2016

### EIXO TEMÁTICO:

Metadados, Acesso e Recuperação de dados e Informação


## UMA ABORDAGEM PARA CRIAÇÃO DE VALOR EM DADOS ABERTOS PARA PEQUENAS E MÉDIAS EMPRESAS UTILIZANDO O ECOSISTEMA R

### *AN APPROACH TO VALUE CREATION IN OPEN DATA FOR SMALL AND MEDIUM-SIZED ENTERPRISES USING "R" ECOSYSTEM*

**Resumo:** Com a proliferação de tecnologias voltadas para Internet, tais como computação em nuvem, dispositivos móveis, mídias sociais e para a internet das coisas, pequenas e médias empresas, apesar de possuírem tamanho menor, menos receita e número pequeno de funcionários, não podem operar com menos informação. Existem muitas oportunidades para estas organizações obterem a capacidade de uma compreensão intuitiva precisa e valiosa a partir dos dados. Como a quantidade de dados disponíveis atualmente na Economia da Informação são muito grandes, órgãos públicos e privados buscam aumentar sua transparência publicando dados relevantes on-line e compartilhando-os com o público. No entanto, são necessárias ferramentas adicionais para a geração de valor potencial nestes dados. Ao lidar com a criação de valor é essencial identificar os principais beneficiários do valor que está sendo criado. Desta forma, esta massiva quantidade de dados acessíveis, nas mais variadas formas e plataformas, incentivam as organizações em obter dados que possam ajudá-las a melhorar sua prestação de serviços. O objetivo da pesquisa, ainda em desenvolvimento, é o de criar aplicativos, utilizando tecnologia aberta, para o tratamento e visualização de informações com vistas a torná-las mais explícitas criando valor em tais dados abertos. No intuito de validar o conceito teórico estabelecido na pesquisa, neste artigo realizamos uma prova de conceito com dados abertos de um órgão do governo federal que apoia e financia investimentos em pequenas e médias empresas. Os resultados obtidos mostraram ser possível elaborar previsões sobre os possíveis desembolsos de investimentos para o próximo ano em cada setor.

**Palavras-chave:** Análise de Dados; Dados Abertos; Linguagem R; Pequenas e Médias Empresas; Grandes Dados.

**Abstract:** With the technologies proliferation for Internet, such as cloud computing, mobile devices, social media and the internet of things, small and medium enterprises, despite their smaller size, less revenue and small number of employees cannot operate with less information. There are many opportunities for these organizations gain the ability of an intuitive, precise and valuable understanding from the data. As the amount of currently available data in Information Economy are very large, public and private agencies seek increase transparency by publishing relevant online data and sharing them with the public. However, additional tools are needed to generate potential value from these data. When dealing with value creation it is essential identify the main beneficiaries from value being created. Thus, this massive amount of accessible data, in diverse forms and platforms, encourage organizations obtain



data that can help them improve their services. The research's objective, still in development, is to create applications using open technology for information treatment and display in order to make them more explicit creating value in such open data. In order to validate the principal research's theoretical concept, this paper conducted a proof of concept with open data from a federal government agency that supports and finances investments in small and medium enterprises. The results proved to be possible to develop predictions about the possible investment expenditures for next year in each sector.

**Keywords:** Analytics; Open Data; R Language; Small and Medium Enterprises; Big Data.



## 1 INTRODUÇÃO

Dada a massiva quantidade de dados disponíveis, nas mais variadas formas e plataformas, os mesmos podem ser definidos como "grandes dados" (big data). Com a proliferação de tecnologias voltadas para Internet, tais como computação em nuvem (cloud computing), dispositivos móveis, mídias sociais e internet das coisas (IoT - internet of things), micros, pequenas e médias empresas (MPME), que possuem tamanho menor, menos receita e número pequeno de funcionários, podem também se beneficiar desta vasta informação disponível e gerar valor para si próprias, afirma Preez (2014).

De fato, um relatório recente da Research and Markets (2015) prevê que o mercado global de grandes dados para pequenas e médias empresas (PME) vai crescer a uma taxa anual composta de 43% até 2019. Pequenas e médias empresas representam a maioria das empresas e são responsáveis por ofertar grande parcela do emprego total no setor privado na maioria das economias. (BECK, 2007).

A medida que a economia do Brasil cresce, crescem também as oportunidades para os empresários de PMEs. O país tem atualmente cerca de 6,3 milhões de pequenas e médias empresas. No Brasil, o Banco Nacional de Desenvolvimento Econômico e Social (BNDES) define as PMEs como as empresas que possuem receita operacional anual de US\$ 38 milhões ou menos. (LEME, 2015).

Ainda segundo LEME (2015), os principais fatores e tendências para as PME brasileiras podem ser resumidos como: (a) as PME contribuem em 20 % do PIB do Brasil; (B) mais de 50 % dos empregos formais do Brasil estão nas PME; (c) no Brasil, a tecnologia tem papel crescente nas PME; e (d) o empreendedorismo é uma escolha para o crescimento na carreira, mais que uma necessidade.

Como a quantidade de dados disponíveis atualmente na Economia da Informação são muito grandes os órgãos públicos e privados buscam aumentar a transparência dos processos e do seu desempenho publicando dados relevantes on-line e compartilhando-os com o público. (The Economist, 2010).

De acordo com o Ubaldi (2013), os dados de governos, disponibilizados em conjuntos legíveis e vinculados se constituem em um novo recurso crítico para abastecer mudanças na criação de valor aos seus diversos usuários. Ao lidar com a criação de valor é essencial identificar os principais beneficiários do valor que está sendo criado, segundo o relatório da



OCDE. (UBALDI, 2013).

Existem muitas oportunidades para as PMEs obterem insights extremamente valiosos e novos a partir de dados proprietários já existentes ou no uso de novos conjuntos de dados. As pequenas e médias empresas têm que pensar em soluções inovadoras para ver oportunidades nos dados internos, bem como de fora de fora de sua organização. Portanto, uma vez que isto se torne também um fator importante para as PMEs, e se estas pequenas corporações estiverem dispostas a investir tempo e dinheiro em uma estratégia de uso de análise de dados para agregar valor ao seu negócio, elas também podem alcançar resultados consideráveis e superar os seus concorrentes.


Desta forma, a proposta deste trabalho, o qual faz parte de uma pesquisa em desenvolvimento, é a de gerar conhecimentos sobre o ambiente de desenvolvimento R para aplicação prática, dirigida às pequenas e médias empresas, demonstrando, em uma abordagem simplificada, a construção de algoritmos flexíveis no ecossistema “R”, para análise de dados abertos do BNDES com referência às PMEs, buscando prever o valor dos desembolsos do banco, para o próximo ano. Como parte do escopo mais geral da pesquisa, este trabalho objetiva também, proporcionar maior familiaridade com os temas "grandes dados", dados abertos, tratamento e visualização gráfica de informação, com vistas a tornar explícito os conceitos de visibilidade e reconhecimento adicional (criação de valor) em tais dados abertos.

## 2. FUNDAMENTAÇÃO TEÓRICA

Pesquisa realizada pelo SAS Institute (2013) mostra que 71% das organizações ainda não têm uma estratégia de análise de dados. A grande maioria das organizações ainda não começou a testar ou mesmo a planejar uma estratégia de "big data". De acordo com essa pesquisa, 76% dos altos executivos das PMEs entrevistados visualizam a utilização de análise de dados como uma oportunidade para uso e melhoria na gestão de suas empresas. (RIJMENAM, 2015).

Big data não é apenas sobre volume e velocidade dos dados e, definitivamente, não é apenas adequado para organizações grandes. Existem amplas oportunidades para as PME obter insights extremamente valiosos e novos a partir de seus dados existentes ou novos conjuntos de dados.

Pelo lado da TI, técnicas de aprendizado de máquina (*machine learning*) são cada vez



mais comuns no mundo dos negócios e de pesquisas para prever vendas e gestão de clientes, ou para ajudar os pesquisadores a compreender e obter insights sobre eventos. (DEJAEGER et. ali., 2012).


Segundo Simon (2013), a aprendizagem de máquina (AM) é um subcampo da ciência da computação que evoluiu a partir do estudo do reconhecimento de padrões e da teoria da aprendizagem computacional em inteligência artificial. A aprendizagem automática explora o estudo e construção de algoritmos que podem aprender e fazer previsões sobre os dados. Tais algoritmos operam através da construção de um modelo de um conjunto exemplo de treinamento de observações de entrada, a fim de fazer previsões baseadas em dados ou decisões expressas como saídas, ao invés de seguir de forma estritamente estática as instruções do programa.

Definido como "análise preditiva", a técnica de aprendizagem automática, de acordo com Simon (2013), se refere ao processo de coleta e processamento de dados e a aplicação posterior de alguma forma de análise matemática nesses dados para obter informações valiosas. Com a automatização dos esforços de coleta de dados, cada vez mais dados estão sendo capturados, tornando a tarefa de extrair padrões interessantes cada vez mais desafiadora.

Embora a abordagem de a aprendizagem de máquina requiera pessoal especializado, atualmente não exige custo muito elevado. De acordo com Mitchell et. ali. (2012), muitas ferramentas de código aberto são gratuitas e o custo do hardware padrão (commodity) se torna cada vez mais barato. O código-fonte aberto é uma parte importante da Web 3.0, porque se encaixa em um dos principais fatores na sua promoção: a colaboração.

De acordo com Markoff (2006), a *Web 3.0* representa a terceira geração da Internet. Nesta nova geração os conteúdos online estarão organizados de forma semântica e muito mais personalizados para cada usuário: os sites, aplicações inteligentes e publicidade serão baseados nas pesquisas e nos comportamentos dos internautas.

O movimento "*open source*" (código aberto em software) obteve um ambiente seguro para a sua consolidação com o surgimento da Internet, assim como também reforçou a necessidade para atender a reformulação maciça dos códigos-fonte de computação. Um princípio fundamental do desenvolvimento de software de código aberto é o da produção pelos pares, no qual produtos como o código-fonte, "*blueprints*", e documentação ficam disponíveis para o público, sem nenhum custo. O movimento de código aberto em software começou como uma resposta às limitações do código proprietário, e, desde então, tem se espalhado por



diferentes campos.

Já os Dados Abertos Governamentais (DAG) podem ser utilizados para ajudar as organizações e os cidadãos a entender melhor as políticas públicas e quanto bem ele executa suas estratégias e torná-lo responsável por ilegalidades ou resultados não alcançados. Isto é particularmente verdadeiro pois uma quantidade considerável destes dados de governo estão se tornando progressivamente mais acessíveis e podem ser usados agregados a informações de outras fontes (por exemplo, informações proprietárias).

Por outro lado, dos governos espera-se que sejam capazes de oferecer facilmente uma gama ampla de dados para promover a tomada de decisões baseada em proeminências. Segundo Ubaldi (2013), os DAG também são vistos como uma fonte importante de crescimento econômico, novas formas de empreendimentos e inovação social.

Os dados são um ativo e os dados que o governo detém os tornam um ativo extremamente valioso. Sendo assim, começa-se a revelar o valor destes dados quando se combinam diferentes fontes de dados relevantes. Portanto, a transparência nos DAG é uma característica importante, mas o valor real dos dados só pode ser adquirido se devidamente explorados pois precisam ser analisados, a fim de se tornarem úteis. Quando se começa a analisar os dados e utilizá-los para a inovação, otimização, previsão e predição, pode-se começar a evoluir em termos de informações transparentes, afirma Manochaan (2011).

De acordo com Smith (2015), a maior disponibilidade de dados fez da ciência de dados crucial para o desenvolvimento, criação e gestão de inovações e novos produtos que são demasiado complexas para os sistemas automatizados (ERPs), especialmente em um mundo onde as preocupações com o sigilo são fundamentais.

Como resultado, as empresas que antes dependiam de plataformas legadas proprietárias para análise estatística estão agora adotando uma nova alternativa, a linguagem de código aberto R. (SMITH, 2015). Em grande parte, devido à sua natureza de código aberto, o R foi rapidamente adotado pelos departamentos de estatística em universidades do mundo todo, atraídos pela sua natureza extensível como uma plataforma para pesquisa acadêmica. Ser livre de custo certamente também desempenhou um papel.

## 2.1. AMBIENTE DE PROGRAMAÇÃO R

Dentre os vários ambientes que oferecem a possibilidade de programação de código aberto para análise de dados, a plataforma R vem se tornando, de forma crescente, um padrão

de fato.

De acordo com Diakopoulos e Cass (2015), das 10 linguagens de programação mais utilizadas, a linguagem R é a que demonstra maior crescimento. A Tabela 1 compara o desempenho das linguagens no ano de 2016 (primeira coluna “Spectrum Ranking”) de acordo com pesquisa da IEEE.

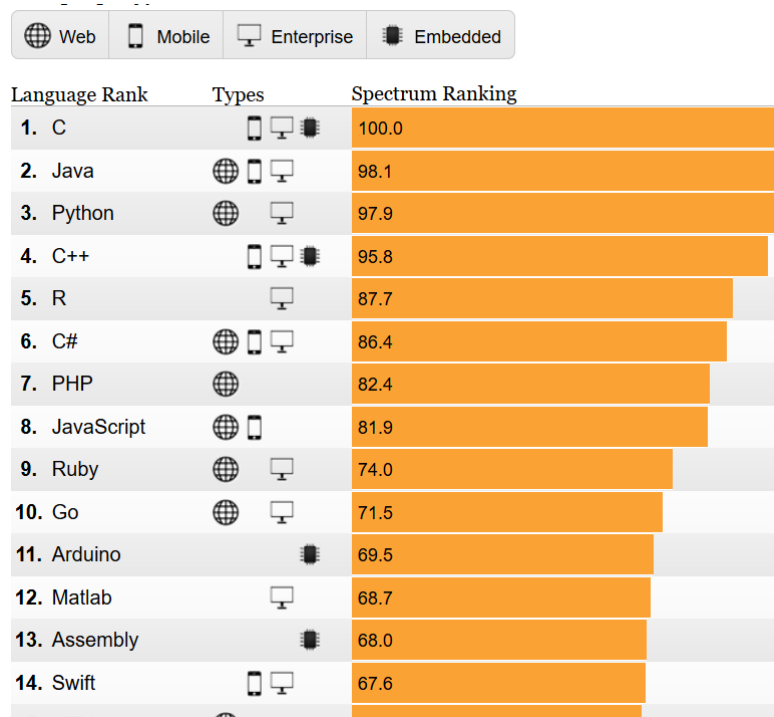



Tabela 1. Linguagens mais utilizadas em TI. Fonte: Diakopoulos e Cass (2016).

Como se pode observar na Tabela 1, as quatro grandes linguagens, C, Java, Python e C++ permanecem no topo. Entretanto, de acordo com Diakopoulos e Cass (2016), o grande salto na classificação foi o da linguagem R, uma linguagem de computação estatística que é útil para análise e visualização de grandes dados, que está em quinto lugar. No ano de 2014, a linguagem R estava em nono lugar e seu movimento reflete a crescente importância dos grandes dados para um grande número de áreas.

O R é uma linguagem e ambiente para computação e gráficos estatísticos, semelhante à linguagem “S” originalmente desenvolvida nos Laboratórios da Bell. É uma solução open source (de código aberto) para análise de dados que é apoiada por uma comunidade de pesquisa muito grande e ativa em todo o mundo. (THE R PROJECT, 2015).

O R é uma parte oficial do projeto GNU da Free Software Foundation e da Fundação R e tem objetivos semelhantes aos outros fundamentos de software de fonte aberta como a





Fundação Apache ou a Fundação GNOME.

## 2.2. PROCESSO DE CIÊNCIA DE DADOS

De acordo com Dhar (2013), Ciência de Dados é um campo interdisciplinar sobre processos e sistemas utilizados para extrair conhecimento ou "insights" dos dados nas formas estruturadas ou não estruturadas. Além disso, segundo Leek (2013), a ciência de dados pode ser considerada uma continuação de alguns dos campos da análise de dados, tais como estatística, aprendizagem de máquina (*machine learning*), mineração de dados (*data mining*) e análise preditiva, e considerada similar a Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases).

A ciência de dados afeta a academia e a pesquisa aplicada em muitos domínios, como por exemplo, tradução automática, reconhecimento de voz, robótica, motores de busca, economia digital, bem como também tem atuação em ciências biológicas, informática médica, cuidados da saúde, ciências sociais e humanidades, afirma Leek (2013). Também influencia fortemente a economia, negócios e finanças. Do ponto de vista empresarial, a ciência de dados é uma parte integrante da inteligência competitiva, um campo emergente que engloba várias atividades, como a análise e mineração de dados de dados. (DHAR, 2013).

Na concepção de ciência de dados, os dados representam os vestígios de processos do mundo real, e exatamente quais traços devem ser coletados são decididos pela nossa coleção de dados ou método de amostragem. Após a coleta e verificação dos dados, utiliza-se um modelo matemático simplificado e conciso para representar toda a população. Este processo de se deslocar do mundo para os dados e, em seguida, a partir dos dados voltar para o mundo, é o campo da inferência estatística.

O modelo básico deste *processo de ciência de dados* utilizado neste artigo será o modelo definido em Schutt e O'Neil (2014) e representado na Figura 1. Em resumo, esse modelo se baseia nas seguintes etapas:

1. Análises estatísticas;
2. Mudança dos dados para outro formato (transformação) para ser usado ou processado: (análise, raspagem e formatação de dados)
3. Visualização (gráficos, sumários, ferramentas, etc.).



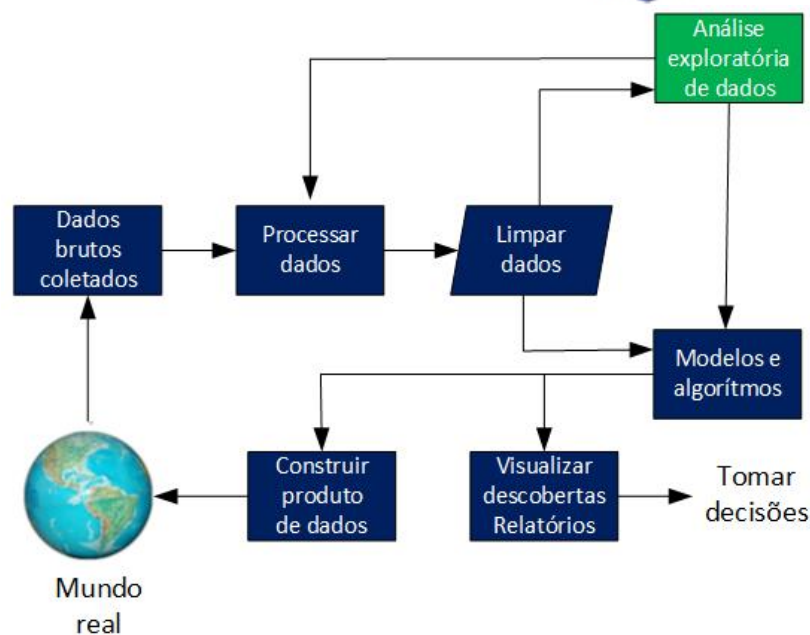


Figura 1. Processo de Ciência de Dados. Fonte: Schutt e O'Neil (2014).

Portanto, segundo Schutt e O'Neil (2014) e seguindo o fluxo da Figura 1, os dados são coletados a partir do mundo real ("dados em brutos") e inicialmente devem ser transformados e “limpos” para análise (data munging). Uma vez que o conjunto de dados está limpo e preparado, técnicas de Análise Exploratória de Dados (AED) são utilizadas para posterior análise. Algoritmos de aprendizagem de máquina (machine learning) são utilizados e, finalmente, as informações são representadas utilizando-se técnicas e ferramentas de visualização de dados.

### 3. METODOLOGIA DE PESQUISA

De acordo com Gil (2008) uma pesquisa é exploratória quando busca estabelecer critérios, métodos e técnicas e visa oferecer informações sobre o objeto desta e orientar a formulação de hipóteses.

A pesquisa tecnológica exploratória oportuniza a obtenção de patentes nacionais e internacionais, a geração de riquezas e a redução da dependência tecnológica, afirmam Barquette e Chaoubah (2007). Novos produtos e processos podem ser originados por impulsos criativos, que a partir de experimentações exploratórias produzem invenções ou inovações. Nas atividades exploratórias concentram-se as principais descobertas científicas, muitas originadas pelo acaso quando da constatação de fenômenos ocorridos durante experimentos em



laboratórios.

Neste sentido, esta pesquisa se configura como exploratória pois visa criar um método e uma técnica para agregar valor a dados abertos desenvolvendo códigos em plataforma aberta, para apoiar a análise e visualização de tais dados disponibilizados nos portais na web. Para tanto, utiliza a linguagem R e seu ecossistema de pacotes além de informações colhidas via Portal Brasileiro de Dados Abertos.

A principal questão de pesquisa do projeto é:

*“pode-se agregar valor a dados abertos nos portais web por meio de códigos em plataforma aberta, apoiando sua análise e visualização?”*

Para responder à questão de pesquisa foram realizadas pesquisas bibliográficas e *sitiográficas* em sites de dados abertos, no Portal Brasileiro de Dados Abertos com o intuito de estabelecer as fontes iniciais de dados (dados brutos).


Para o desenvolvimento das ferramentas para compreensão e análise dos dados abertos foram realizadas pesquisas iniciais nos portais rOpenGov (<http://ropengov.github.io/>) e Data.Gov (<https://www.data.gov/>) além dos sites Bioconductor e rOpenSci dentre outros, com o objetivo de utilizar a experiência e a documentação de lições aprendidas.

#### **4. APLICAÇÃO DO PROCESSO DE CIÊNCIA DE DADOS**

Buscando elucidar provar o conceito estabelecido nos tópicos desenvolvidos nos itens anteriores deste artigo, realizaremos nos itens a seguir uma *prova de conceito* usando como modelo básico o *processo de ciência de dados* definido por Schutt e O'Neil (2014) e representado na Figura 1.

De acordo com o modelo da Figura 1, a primeira etapa importante antes de resolver um problema é definir exatamente qual é o problema (do Mundo Real) se espera resolver. No caso de um processo de ciência de dados é preciso ser capaz de traduzir as questões sobre os dados em algo acionável (Schutt e O'Neil, 2014).

Neste caso especificamente, buscamos resolver um problema para as MPEs, procurando determinar quais seriam os possíveis desembolsos de investimentos para o próximo ano neste setor do Banco Nacional de Desenvolvimento Econômico e Social (BNDES, 2016)), principal instrumento do Governo Federal brasileiro para o financiamento de longo prazo e investimento em todos os segmentos da economia brasileira. Os dados dos anos anteriores (2006 a 2015)



estão disponíveis no sítio do BNDES (2016)<sup>1</sup>.

A questão principal será realizar uma *previsão* se o BNDES (2016) continuará a investir nesse segmento nos mesmos patamares dos anos anteriores. O importante, no final desta primeira etapa, é obter todas as informações e o contexto necessários para resolver o problema.

Após definir o problema, é preciso obter os dados necessários e construir os conhecimentos necessários para transformar o problema em torno de uma solução. Esta etapa do processo envolve raciocinar quais os dados são necessários e encontrar maneiras de obtê-los, quer se trate de consultar bancos de dados internos, ou a aquisição de conjuntos de dados externos.

No exemplo específico, foram obtidos os “dados brutos” no formato Excell (.xlsx). A Figura 2 mostra o trecho do código fonte em R para a carga do arquivo no ambiente da linguagem.

```
# download, leitura e preparação do dataset
require(gdata)
require(rJava)
require(xlsx)
cnae<-read.xls("Int2_1D_a_setorCNAE_MPME.xlsx", header=TRUE, row.names=1)
```

Figura 2. Código fonte em R para a carga do arquivo. Fonte: Elaborada pelo autor.

Após a carga dos dados brutos, é necessário avaliá-los antes de qualquer análise. Muitas vezes, os dados podem estar bastante confusos, especialmente se não foram coletados e armazenados com cuidado, contendo erros que poderão corromper a análise, tais como: valores definidos como nulos embora sejam realmente zero; valores duplicados; e valores ausentes (*missing values*). Cabe ao cientista de dados avaliar e verificar para se certificar que os dados possuem informações precisas. Esta etapa, na Figura 1, correspondem aos processos de transformação (processamento) e limpeza dos dados.

O próximo passo é a realização de uma análise exploratória dados. A análise exploratória de dados (AED) é o primeiro passo a ser dado para a construção de um modelo. A AED é uma parte crítica do processo de ciência de dados e também representa uma filosofia ou modo de realizar processos estatísticos praticados por uma estirpe de estatísticos provenientes da tradição dos Laboratórios Bell Labs, afirmam Schutt e O’Neil (2014).

---

<sup>1</sup> A planilha original, no formato PDF está disponível em:

<[http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes\\_pt/Galerias/Arquivos/empresa/estatisticas/Int2\\_1D\\_a\\_setorCNAE\\_MPME.pdf](http://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/empresa/estatisticas/Int2_1D_a_setorCNAE_MPME.pdf)>. Acessado em: 13/09/2016 17:02:37.

Os gráficos iniciais da análise exploratória dos dados no exemplo estão representados nos gráficos e histograma da Figura 3.

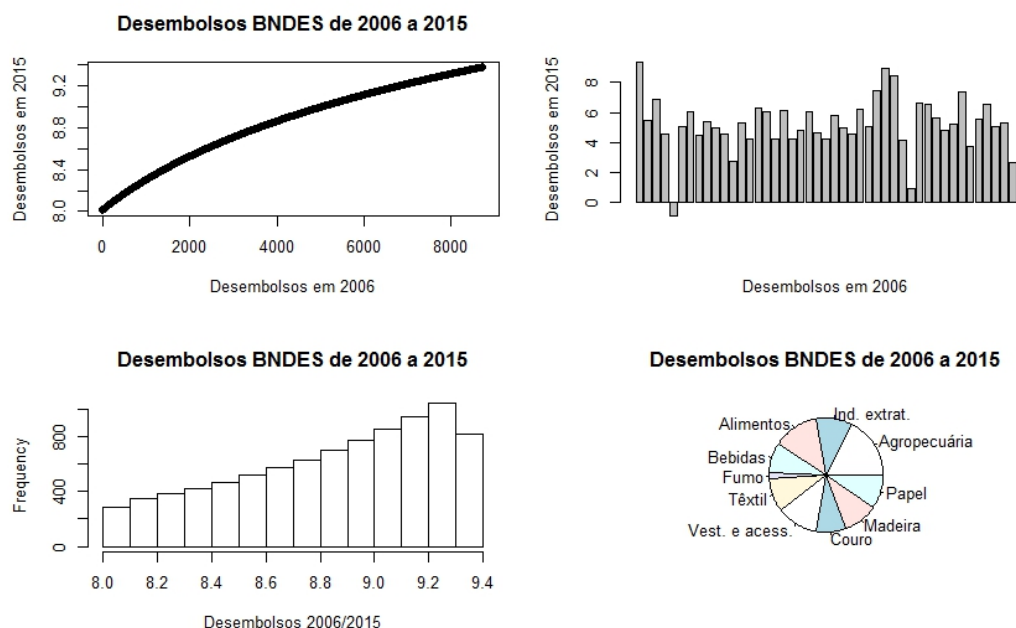


Figura 3. Análise exploratória dos dados. Fonte: Elaborada pelo autor.

A etapa seguinte à análise exploratória é o da *construção do modelo*. No nosso exemplo, buscamos criar um *modelo preditivo* utilizando os dados já existentes nos últimos anos de desembolsos do BNDES (2016) para pequenas e médias empresas de todas as regiões do Brasil e de todas as áreas de negócios.

Nesta etapa, construiu-se um modelo de regressão para tentar prever valores dos desembolsos baseados em outros atributos dos valores (anos anteriores), utilizando o modelo de regressão linear múltipla.

A abordagem de modelagem de regressão linear geralmente consiste de uma variável de resposta, a qual procuramos prever e diversas variáveis de entrada. O modelo também assume que existe uma relação linear entre as variáveis de entrada e nossa variável de resposta.

Na Figura 4 podemos ver os detalhes do modelo utilizando o comando *summary* (resumo) da linguagem R na variável *initial\_model*, o que nos dá informações detalhadas sobre o modelo, os coeficientes das múltiplas variáveis em diferentes métricas.

```
#Call:
#lm(formula = X2015 ~ . - Tipo, data = cnae)
#
#Residuals:
#  Min   1Q Median   3Q   Max
#-96.45 -33.86 -11.11  23.78 170.98
#
#Coefficients:
#      Estimate Std. Error t value Pr(>|t|)
#(Intercept)  16.9865    14.9337   1.137 0.263074
#X2006        -0.3976     0.3109  -1.279 0.209290
#X2007         1.6200     0.3841   4.217 0.000166 ***
#X2008        -0.3114     0.2267  -1.374 0.178313
#X2009         0.7708     0.3664   2.104 0.042660 *
#X2010        -1.5935     0.1318 -12.087 4.77e-14 ***
#X2011         0.6977     0.1841   3.789 0.000573 ***
#X2012        -0.3555     0.1583  -2.246 0.031132 *
#X2013         0.5806     0.1361   4.266 0.000144 ***
#X2014         0.3748     0.1008   3.717 0.000702 ***
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 65.13 on 35 degrees of freedom
#Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9991
#F-statistic: 5213 on 9 and 35 DF, p-value: < 2.2e-16
```

Figura 4. Resumo do modelo. Fonte: Elaborada pelo autor.

Observa-se que o valor de  $R^2$  ajustado (*Adjusted R-squared*) é 0,9991, o que indica que 99,91% de variação da variável resposta (X2015) é explicada pelas variáveis de entrada. Quanto maior este valor, melhor será o modelo porque este irá explicar a maior parte da variabilidade observada na variável de resposta, a qual se busca prever.


Finalmente, na Figura 5 mostramos as 10 primeiras áreas de negócios das pequenas e médias empresas e os valores previstos de desembolsos pelo BNDES (2016) para o próximo ano.

```
> list(predict(initial_model, cnae[1:10,]))
[[1]]
Agropecuária      Indústria extrativa      Produtos alimentícios
11755.96           279.82                956.63
Bebidas           Fumo                  Têxtil
122.92            30.50                70.77
Confec.. vestuário e acessórios  Couro. artefato e calçado  Madeira
419.33            100.31043            257.85
Celulose e papel
78.69
```

Figura 5. Previsão calculada dos desembolsos. Fonte: Elaborada pelo autor.

#### 4 CONSIDERAÇÕES PARCIAIS E POSSÍVEIS APLICAÇÕES DO ESTUDO

Este artigo tem como expectativa incitar o debate no Brasil referente ao uso dos Dados



Governamentais Abertos (DGA) disponibilizados pelos governos federal, estadual e municipal, bem como de organismos da sociedade civil e da necessidade de facilitar a visualização e interpretação de tais dados.

Com esta questão de pesquisa em perspectiva, buscou-se introduzir o ambiente de programação de *código aberto* R e seu ecossistema com o intuito de demonstrar a facilidade do uso de tal ambiente.

O artigo buscou também discutir e demonstrar uma abordagem sistemática (processo de ciência de dados) para a criação de aplicativos inteligentes de forma a permitir que equipes de cientistas de dados colaborarem no ciclo de vida de atividades necessárias para transformar esses aplicativos em produtos.

Realizou-se uma prova de conceito, utilizando dados do BNDES (2016) para demonstrar a viabilidade do modelo.

Finalmente, buscou-se com este trabalho lançar as bases para a possibilidade de criar no País a sua própria comunidade de colaboradores no ecossistema R, o que sem dúvida, tornaria mais rápido o desenvolvimento de uma cultura científica de tratamento de dados abertos e suas ferramentas estatísticas e de modelagem.

Como especificado no início do trabalho, este artigo faz parte de um projeto de pesquisa mais amplo, que é o de desenvolver e gerar aplicativos de fácil reprodução e utilização, utilizando tecnologia aberta, com foco em ferramentas para tratamento e visualização de dados com o intuito de auxiliar organizações e cidadãos que queiram utilizar dados abertos de forma eficaz em seus projetos e até mesmo a ampliar a consciência do potencial dos “grandes dados” para o seu trabalho.

## REFERÊNCIAS

- Advanced ‘Big Data’ Analytics with R and Hadoop, White Paper. **Revolution Analytics**. Washington, USA, 2011. 15 p.
- BARQUETTE, S, CHAOUBAH, A., **Pesquisa de marketing**. São Paulo: Saraiva, 2007.
- BECK, T., **Financing Constraints of SMEs in Developing Countries: Evidence, Determinants and Solutions**, Other publications TiSEM, Tilburg University, School of Economics and Management, AB Tilburg, The Netherlands, 2007. 35 p.
- CARVER, L., “R” You Ready, **QuickRead Featured**, Oct. 28, 2015. Disponível em < <http://quickreadbuzz.com/2015/10/28/r-you-ready/> >. Acesso em: 08/09/2016 14:42.
- DHAR, V., **Data Science and Prediction**", Communications of the ACM, 56 (12): 64, 2013. doi:10.1145/2500499.



Data, data everywhere, **The Economist**, February 25rd, 2010. Disponível em < <http://www.economist.com/node/15557443>>. Acesso em: 20/03/2016.

DEJAEGER, K., LOUIS, P., VANDEN BROUCKE, S., EEROLA, T., GOEDHUYS, L., BAESSENS, B., Beyond the Hype: Cloud Computing in Analytics, 2012. DOI: <http://dx.doi.org/10.2139/ssrn.2165720>.

DIAKOPOULOS, N., CASS, S., Interactive: The Top Programming Languages, IEEE Spectrum, 2016. Disponível em: < <http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016>>. Acesso em: 08/09/2016 10:55.

Estatísticas operacionais para download, **BNDES**, 2016. < [http://www.bndes.gov.br/SiteBNDES/bndes/bndes\\_pt/Institucional/BNDES\\_Transparente/Estatisticas\\_Operacionais/estatisticas\\_download.html](http://www.bndes.gov.br/SiteBNDES/bndes/bndes_pt/Institucional/BNDES_Transparente/Estatisticas_Operacionais/estatisticas_download.html)>. Acesso em: 08/09/2016 14:49.

**Gil, A. C. Como elaborar projetos de pesquisa.** São Paulo, Atlas, 2008.

Global Software-defined Storage (SDS) Market 2015-2019, **Research and Markets**, Dublin, Feb. 25, 2015. 77 p.

KABACOFF, R. I., **R in Action: Data analysis and graphics with R**, Second Edition, Manning Publications Co., New York, USA, 2015. 608 p.

KRILL, P., Why R? The Pros and Cons of the R Language, **InfoWorld**, Jun. 30, 2015. Disponível em < <http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html>>. Acesso em: 19/02/2016.

LEME, L., São Paulo 2014 Blog: Six Facts about SMEs in Brazil, AS/COA - **Americas Society Council of the Americas Publications**, New York, USA, Wednesday, March 12, 2014. Disponível em < <http://www.as-coa.org/blogs/s%C3%A3o-paulo-2014-blog-six-facts-about-smes-brazil>>. Acesso em: 19/02/2016.

LEEK, J., The key word in "Data Science" is not Data, it is Science, **Simply Statistics**, 12-Dec. 2013. Disponível em <<http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/>>. Acesso em: 05/09/2016.

MARKOFF, J., Entrepreneurs See a Web Guided by Common Sense, **The New York Times**, Nov. 12, 2006. Disponível em < [http://www.nytimes.com/2006/11/12/business/12web.html?\\_r=0](http://www.nytimes.com/2006/11/12/business/12web.html?_r=0)>. Acesso em: 19/02/2016.

MITCHELL, M., TURNER, J., CORBETT, R. S., TASNER, M., **Measure the Impact of Online Marketing** (Collection), Financial Times Press, New Jersey, USA, 2012. 329 p.

PREEZ, D. D. Big Data for Small Business, **Raconteur Media Ltd.**, September 9, 2014. Disponível em < <http://raconteur.net/technology/big-data-for-small-business>>. Acesso em: 19/02/2016.

**R: A Language and Environment for Statistical Computing, The R Project**, R Foundation for Statistical Computing. Manual de Programação. Vienna, Austria, 2015. 3501 p.

RIJMENAM, M. V., Also SMEs Can Achieve Remarkable Results with Big Data, **Datafloq**, May 2015. Disponível em < <https://datafloq.com/read/also-smes-can-achieve-remarkable-results-with-big-/192>>. Acesso em: 23/03/2016.

SCHUTT, R., O'NEIL, C. **Doing Data Science: Straight Talk from the Frontline**, O'Reilly Media, California, USA 2014. 408 p.

**SAS Overview and 2013 Annual Report**, SAS Institute., North Caroline, USA, 2013. 9 p.

SIMON, P., **Too Big to Ignore: The Business Case for Big Data**, John Wiley & Sons, New Jersey, USA, 2013. 256 p.

SMITH, D., **Why now is the time to learn R?** Open source.com, 2015. Disponível em < <https://opensource.com/business/14/12/r-open-source-language-data-science>>. Acesso em: 08/09/2016 11:03.

UBALDI, B., *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*, OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris. DOI: <http://dx.doi.org/10.1787/5k46bj4f03s7-en>.