Uso de ferramentas de *Big Data* no auxílio do controle de epidemias de dengue no Brasil

Belo Horizonte, janeiro de 2016.
Érico Martins e Silva – erico@datafoco.com.br
github.com/ericomartins – kaggle.com/ericodmg

Trabalho realizado para o curso "*Big Data* em Saúde no Brasil", ministrado pelo Prof. Alexandre Dias Porto Chiavegatto Filho da Universidade de São Paulo, Faculdade de Saúde Pública, na Plataforma Coursera de cursos online. (Maiores informações: https://www.coursera.org/course/bigdatabrasil)

1 – Introdução

A saúde no Brasil atravessa uma verdadeira cruzada no combate ao mosquito *Aedes aegypti*. Parte deste esforço é composto da coleta de dados em campo dos mais diversos aspectos relativos ao mosquito e a dados epidemiológicos.

As ferramentas de *Big Data* oferecem uma verdadeira revolução no tratamento e análise de dados em cenários complexos, como os epidemiológicos.

O uso destas ferramentas nos esforços de mobilização e controle da epidemia de dengue é uma solução que pode oferecer um grande ganho no potencial de análise do problema, utilizando informações já existentes, com uma excelente relação custo/benefício.

2 - Objetivos

O objetivo principal deste trabalho é incentivar o uso das técnicas de *Big Data* no atual cenário de mobilização e controle da epidemia de dengue no Brasil.

Esse trabalho visa ainda como objetivos específicos apresentar um exemplo básico do uso de ferramentas de Big Data para predição de ciclo de casos da dengue e demonstrar o potencial dessas ferramentas na predição de dados.

3 – Contextualização

O acesso a dados públicos, sobretudo governamentais, tem experimentado uma evolução significativa impulsionado pelos avanços tecnológicos contemporâneos[1] somados à pressão crescente de transparência na administração pública.[2]

A Saúde é um dos campos onde este efeito se torna mais observado. Especificamente no Brasil, a disponibilização de dados da área da saúde tem aumentado indiscutivelmente, despontando inclusive em escala mundial.[3] E essa evolução traz com ela o desafio da demanda por novas formas de análise de dados, disponíveis em grandes volumes, de alta complexidade e muitas vezes desestruturados.

A saúde no Brasil atravessa no momento uma verdadeira cruzada no combate ao mosquito *Aedes aegypti*, justificada pela volta do surto de casos de dengue que havia sido amenizado em 2014, pelo crescimento de febre de chikungunya e febre pelo vírus Zika além do recente aumento da incidência de casos de microcefalia sobretudo no nordeste.[4]

Esta movimentação nacional no combate à dengue é um bom exemplo da captação de um volume extraordinário de dados. As ações de prevenção e combate a dengue mobilizam um número enorme de profissionais, entre entre eles agentes comunitários de saúde, agentes de combate a endemias, profissionais de saúde, educação e até mesmo reforços vindos das Forças Armadas e Defesa Civil.[5] Estas ações não só resultam na captação de dados relacionados à

endemia, como também demandam cada vez mais de inteligência na compreensão destes dados, para garantir sua efetividade e orientar sua logística.

Uma solução que desponta como relevante para o processamento eficiente destes dados é o uso das técnicas de *Big Data*, que abrem novas fronteiras na utilização de massas de dados e na sua análise, prometendo uma verdadeira revolução em termos desempenho, resultado obtido e relação custo/benefício.

Trata-se de um conjunto de tendências, metodologias e ferramentas tecnológicas que trazem uma nova abordagem para o tratamento e processamento de grandes conjuntos de dados para fins de entendimento e tomada de decisões.[6]

Entre as técnicas relacionadas ao *Big Data*, destacamos neste trabalho o uso de *Machine Learning* (Aprendizado de Máquina) para a criação de modelos preditivos e simuladores de efeitos complexos, que tem alta aderência ao problema da análise de dados epidemiológicos, como o que enfrentamos com o *Aedes aegypti* neste momento no Brasil.

4 – Definição do problema

Para exemplificar o potencial do uso de *Big Data* no cenário de combate à dengue no Brasil, desenvolvemos uma pequena aplicação de *Machine Learning* como demonstração, com o objetivo de tornar mais claro o modelo de trabalho típico das técnicas de *Big Data*.

Os dados foram obtidos diretamente na área de indicadores e dados estatísticos do portal da Prefeitura do Rio de Janeiro[7]. Eles são compostos apenas pelo número de casos de dengue por mês, área de planejamento, regiões administrativas e bairros do município do Rio de Janeiro e população do bairro no censo IBGE 2010, no período do ano 2000 à 2015. Como informação complementar foi consultada no mesmo portal a área em hectares de cada bairro citado.

A proposta foi criar, usando apenas estes dados básicos de incidência de casos de dengue, um algoritmo de predição da incidência no próximo ano, a partir dos dados obtidos até o final do ano corrente, sem o uso de qualquer outra informação.

5 – Recursos utilizados

Os dados foram copiados diretamente do portal da Prefeitura do Rio de Janeiro a partir da seguinte estrutura: [fig01]

RIO PREFEITURA
SAÚDE

Subsecretaria de Promoção da Saúde, Atenção Primária e Vigilância em Saúde
Superintendência de Vigilância em Saúde
Coordenação de Vigilância Epidemiológica
Gerência de Vigilância de Doenças e Agravos

Número de Casos de Dengue* por Mês, Áreas de Planejamento, Regiões Administrativas e Bairros
Município do Rio de Janeiro, 2011

Dados Atualizados em 12/12/2013

Área Programática, Regiões Administrativas e Bairros	Pop. Censo 2010	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
Total	6.320.446	1.656	6.132	15.459	25.577	15.958	4.471	1.190	587	657	1.053	1.733	3.180	77.653
Ignorado**	0	19	120	229	323	150	39	12	2	3	11	18	57	983
Área de Programática 1.0	297.976	189	666	1.059	1.492	727	88	48	18	25	39	88	144	4.583
I Portuária	48.664	19	80	125	203	117	6	11	4	2	2	7	13	589
Saúde	2.749	2	10	15	22	16	0	2	1	1	0	3	1	73
Gamboa	13.108	4	12	17	45	20	1	1	1	0	0	0	3	104
Santo Cristo	12.330	9	18	57	66	42	4	3	2	1	0	1	6	209
Caju	20.477	4	40	36	70	39	1	5	0	0	2	3	3	203
II Centro	41.142	35	144	226	287	128	13	5	1	0	6	17	36	898

[http://www.rio.rj.gov.br/dlstatic/10112/4536353/4115710/denagueconfirmados_mes2011.htm]

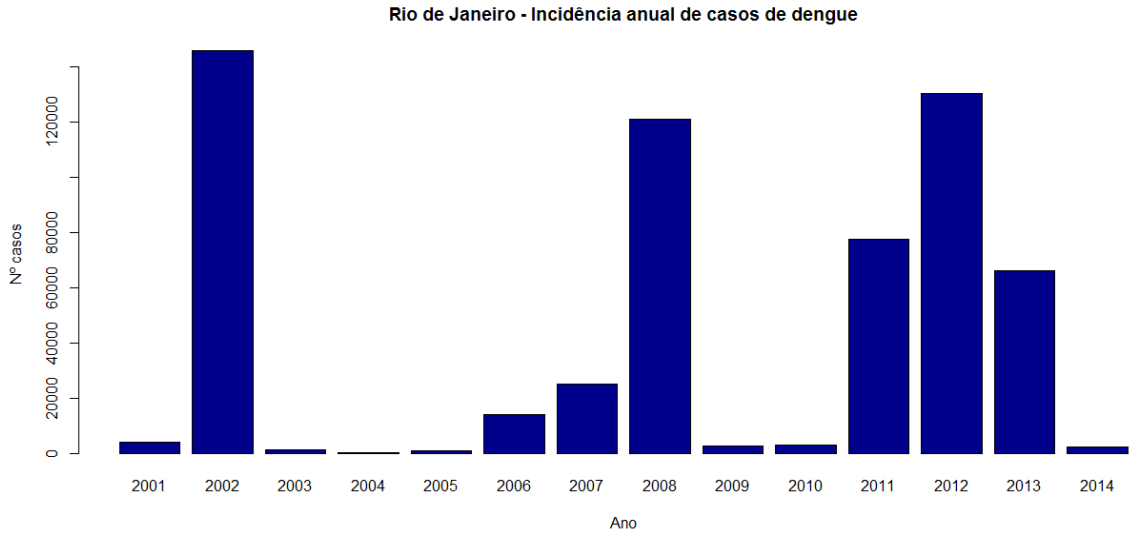
Para a criação do algoritmo optamos por utilizar a linguagem R, pelos seus recursos de implementação de análise de dados e pela tendência mundial de padronizar algoritmos de análise através deste ambiente.

Como mecanismo principal de predição utilizamos um modelo de *Gradient Tree Boosting*, implementado através do pacote *XGBoost - eXtreme Gradient Boosting*, por seu reconhecido desempenho neste tipo de solução.[8]

6 – Método

Como ponto de partida consideramos os dados disponíveis ao final de 2014 com a intenção de prever a curva de incidência do ano de 2015. Para isto vamos criar o algoritmo de predição utilizando somente dados existentes naquele momento. Depois de criado o algoritmo, vamos aplicar o mesmo na predição do ciclo de 2015, e dimensionar a precisão e os erros obtidos.

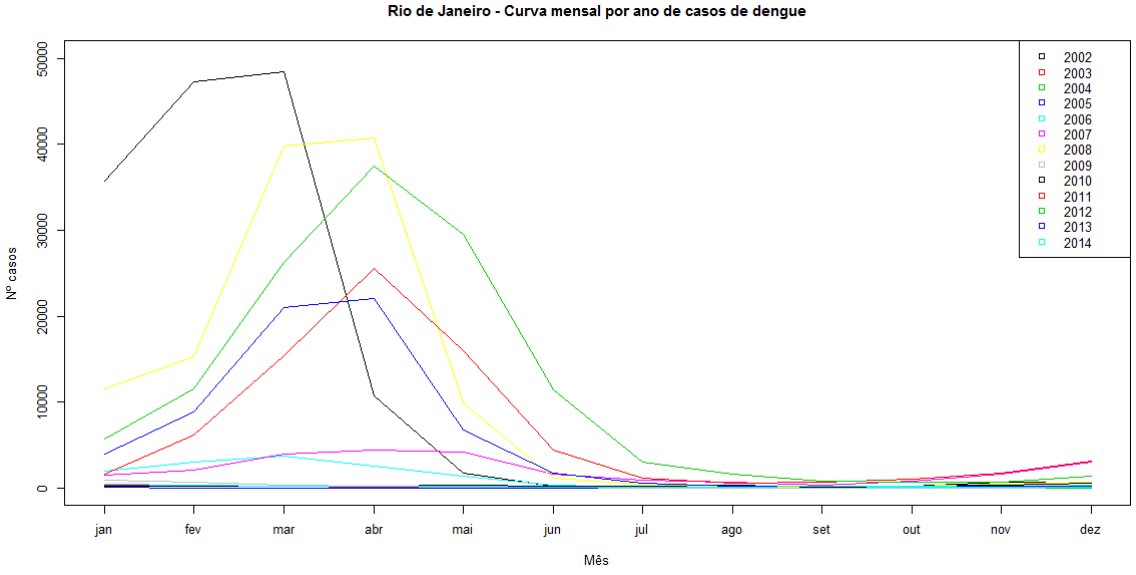
Abaixo apresentamos o total de incidência de casos em cada ano entre 2001 e 2014: [graf01]
(O código que gera todos os gráficos constantes deste documento se encontra no [anexo02])



2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
4.375	145.779	1.516	554	985	14.117	25.332	120.917	2.727	3.268	77.653	130.412	66.278	2.649

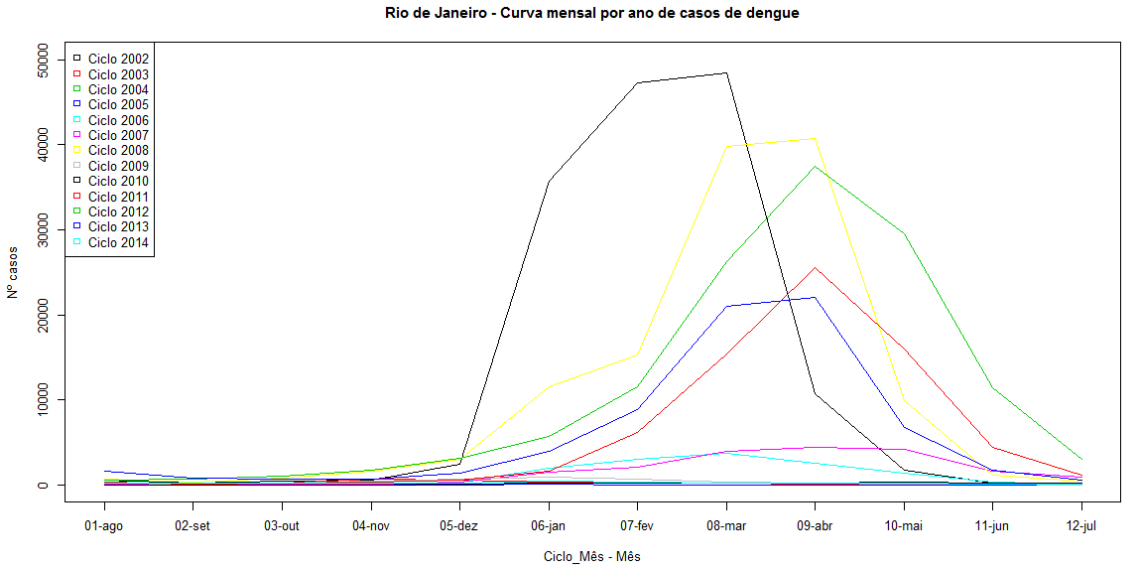
Observamos que o total anual varia entre 554 em 2004 até mais de 145.000 em 2002, sem apresentar uma evolução anual consistente.

Quando analisamos a curva mensal de cada ano, constatamos que apesar dos diferentes totais anuais, existe uma coincidência nos picos de incidência, concentrados entre os meses de fevereiro e abril, com exceção o ano de 2002 que teve um ciclo atípico adiantado : [graf02]



Esta distribuição anual dos dados não é conveniente para a análise do ciclo de dengue, visto que o mesmo apresenta seu início por volta do mês de agosto ou setembro. Por isso, usamos o conceito de "Ciclo", que se inicia no mês de agosto do ano anterior e termina em Julho do ano em que ocorre o pico. Assim, o "Ciclo 2013" irá de agosto de 2012 até julho de 2013, e tem seu pico de incidência por volta de março e abril de 2013. O mês 08/2012 corresponde ao mês 1 do ciclo 2013, o mês 12/2012 corresponde ao mês 5 do ciclo 2013, e o mês 07/2013 corresponde ao mês 12 do ciclo 2013.

A figura abaixo mostra os dados reorganizados nos seus respectivos ciclos: [graf03]



Como estamos considerando os dados até dezembro de 2014 como existentes, já temos do nosso ciclo de previsão (Ciclo 2015) os 5 primeiros meses. Faremos a previsão então dos meses 6 a 12 do Ciclo 2015. (janeiro a julho de 2015). Para facilitar esta tarefa, dividimos cada ciclo em duas partes:

Referência – mês 1 a 5 do ciclo – agosto a dezembro do ano anterior ao pico"
Predição – mês 6 a 12 do ciclo – janeiro a julho do ano da incidência do pico"

A estratégia será utilizar os dados históricos que contém o comportamento das curvas dos ciclos anteriores mais o comportamento inicial do ciclo atual, de agosto a dezembro de 2014, para prever a parte mais significativa da epidemia que ocorrerá de janeiro a julho de 2015.

Visto que precisamos de dados do ano anterior para compor cada ciclo e ainda que os dados de 2000 estão em formato incompatível, serão usados os ciclos de 2002 a 2014, e previsto o ciclo de 2015.

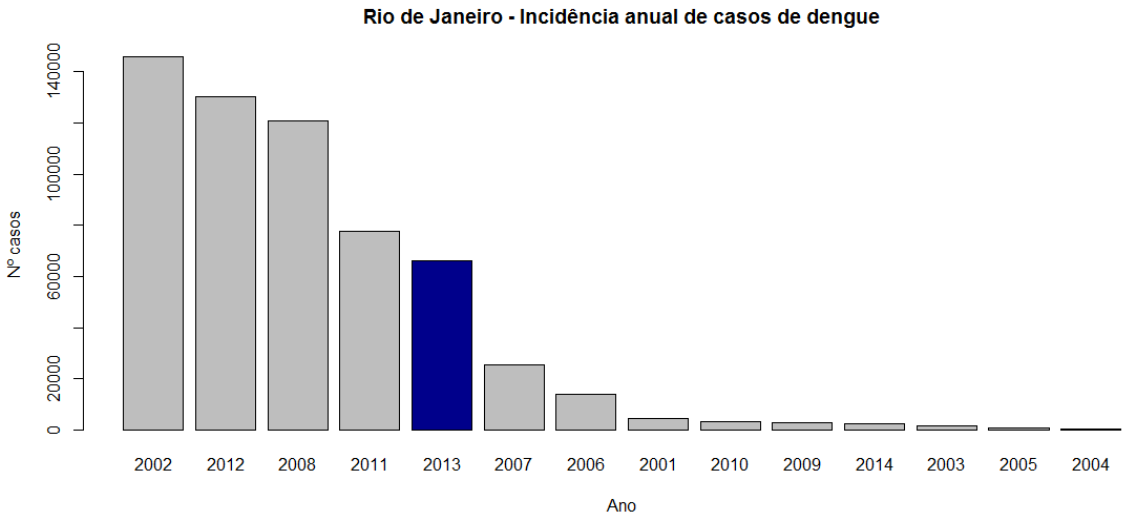
7 – Descrição do processo

Os dados obtidos no site foram planilhados manualmente e após organizados foram salvos no seguinte arquivo no formato csv, detalhado abaixo: [anexo01]

Arquivo data_casos.csv				
caso_ano	**caso_anomes**	**caso_mes**	**qtd_casos**	**bairro**
2002	200201	1	452	Copacabana
Ano do caso	Ano-Mês do caso	Mês do caso	Quantidade de casos no mês, no bairro	Nome do bairro
regiao	**zona**	**area**	**area_grupo**	**bairro_ha**
V RA - Copacabana	Zona Sul	AP 2.1	AG 2	410,09
Região administrativa	Zona à que o bairro pertence	Área programática	Agrupamento criado para áreas	Área do bairro em hectares

bairro_pop2010	ciclo_mes	ciclo	ciclo_parte	ciclo_num
146392	6	ciclo2002	predicao	2002
População no censo IBGE 2010	Mês do ciclo (Inicio em agosto)	Ciclo de análise	Parte do ciclo em que ocorreu o caso	Ciclo de análise em número puro

Será criado um modelo preditivo e durante este processo precisamos testar a qualidade da nossa previsão. Precisamos então definir um ano como sendo nossa referência para teste e refinamento deste nosso modelo preditivo. Visualizamos novamente o total de casos por ano, ordenado agora por total de casos: [graf04]



Vamos definir como nosso ano de teste o ano de 2013, por estar próximo ao ano mediano e por marcar um ponto de inflexão do volume de dados.
O código em R utilizado para rodar os processos criados se encontra no [anexo 3], e explicaremos cada parte relevante do mesmo a seguir.

Processo 1 – Carga:

```
ciclo_teste <- 2013
ciclo_alvo <- 2015
mes_inicio <- 8
ciclo_mes_prev   <- 5
casos <- read.csv("Demo\\anexo01.csv", encoding="UTF-8")
```

O processo 1 inicialmente seta algumas variáveis com os parâmetros para a execução do código, a saber:

ciclo_teste: 2013 - Ciclo escolhido como ciclo de teste.
ciclo_alvo: 2015 - Ciclo que é o alvo para a predição final.
mes_inicio: 8 - Mês do ano que marca o início de um ciclo.
ciclo_mes_ref: 5 - Ciclo_Mês que marca o último mês considerado como parte de referência do ciclo.

Em seguida são lidos os dados do arquivo anexo01.csv e carregados em um *dataframe* denominado **casos**. Ao final do processo 1 vemos o *dataframe* **casos** com o seguinte exemplo de conteúdo:

nrow(casos): [1] 26880
head(casos[casos$bairro == "Copacabana" & casos$caso_anomes == 200201,],1):

caso_ano	caso_anomes	caso_mes	qtd_casos	bairro	regiao	zona	area
2002	200201	1	452	Copacabana	V RA – Copacabana	Zona Sul	AP 2.1

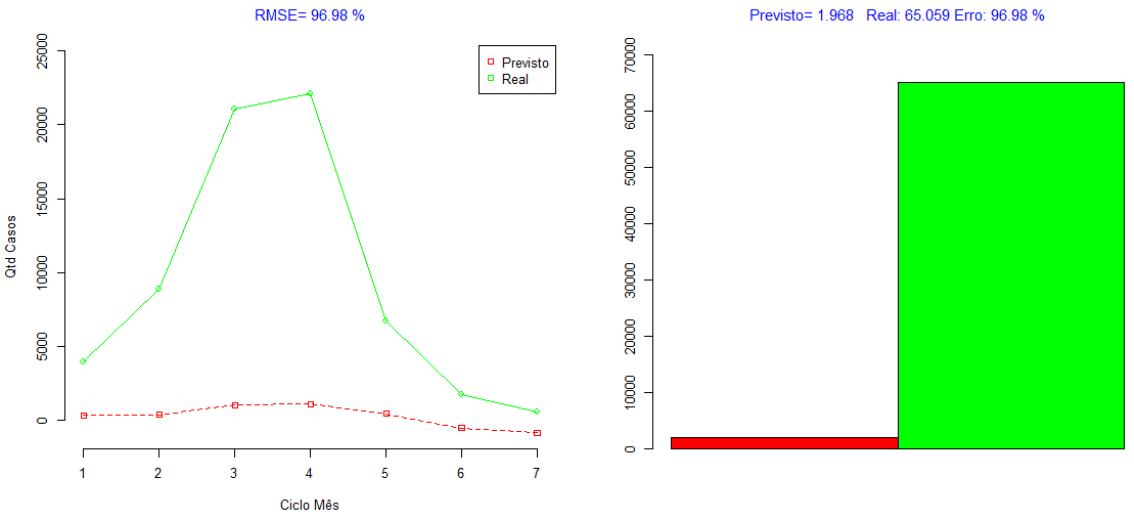
area_grupo	bairro_ha	bairro_pop2010	ciclo_mes	ciclo	ciclo_parte	ciclo_num
AG 2	410.09	146392	6	ciclo2002	predicao	2002

Processo 2 – Teste inicial do modelo:

```
train02 <- subset(casos, ciclo_num != ciclo_alvo | (ciclo_num == ciclo_alvo & ciclo_mes <= ciclo_mes_prev))
train02 <- subset(train02, ciclo_num != ciclo_teste | (ciclo_num == ciclo_teste & ciclo_mes <= ciclo_mes_prev))
train02 <- subset(train02, ciclo_parte == "predicao")
train02 <- train02[order(train02$qtd_casos,train02$bairro,train02$caso_anomes),]
```

```
test02  <- subset(casos, ciclo_num == ciclo_teste & ciclo_mes > ciclo_mes_prev)
modelopred02 <- criamodelo(train02 ,"qtd_casos")
predicao02 <- criapredicao(modelopred02,test02)
```

Este processo passa para o *XGBoost* os dados históricos e verifica inicialmente a qualidade do modelo preditivo obtido. Os dados deverão conter apenas os dados disponíveis no fim de 2014. Usamos as variáveis **ciclo_teste** e **ciclo_alvo** para retirar os dados de predição dos ciclos 2013 e 2015 para que nosso modelo preditivo não seja influenciado por eles, o que inviabilizaria nosso propósito. Foram criados os *dataframes* **train02** e **test02** que contém os dados históricos, sendo que o **train02** tem os dados de todos os ciclos, exceto os dados de previsão dos ciclos de 2013 e de 2015. Já o **test02** possui os meses de predição do ciclo 2013, do ciclo_mes 06 ao 12. O *dataframe* **train02** foi passado para o *XGBoost*, que construiu o modelo preditivo baseado neles, e armazenou esse modelo no objeto **modelopred02**. Em seguida, os dados do **test02** foram aplicados à esse modelo, e foi feita uma predição da variável **qtd_casos** para cada cada bairro/mês nele existente. O resultado totalizado por mês desta predição é apresentado ao final do processo: [fig02]



Usamos para a mensuração do erro da predição o RMSE (*Root Mean Squared Error*).[9] Esta medida soma os erros absolutos de cada unidade mínima de previsão, no nosso caso bairro/mês, e calcula sua média geral. O resultado desta primeira predição não foi satisfatório. Obtivemos um RMSE de 96,98 % nesta primeira validação, e de um real de 65.059 casos, nosso modelo previu apenas 1.968.

Processo 3 – Features derivadas:

```
casos03 <- criafeaturesderivadas(casos,ciclo_teste,ciclo_alvo)
```

Agora precisamos melhorar nosso modelo preditivo, para isto temos de criar novas *features* que enriquecerão a capacidade do *XGBoost* de entendimento do efeito real. Estas *features* serão novas colunas no nosso *dataframe*, que representam efeitos candidatos a influenciadores no processo a ser previsto, no nosso exemplo número de casos de dengue. Através de ciclos iterativos de tentativa, validação e correção de erro, chegamos a um resultado com as seguintes novas *features*, que foram acrescentadas a uma cópia do *dataframe* **casos**, chamada **casos03,** através da função **criafeaturesderivadas**:

totultmesref: Soma dos casos no último mês do período de referência por bairro, de todos os ciclos
desviopadraoref: Desvio padrão da parte referência por ciclo
fatorpred_ultref: Razão da média entre cada mês de predição e o último mês de referência, por bairro
fatorpred_mesant: Razão da média entre cada mês de predição e o mês anterior a ele, por bairro, de todos os ciclos
ciclomenorerro: Total de casos de cada mês de predição do bairro que comparando o período de referência com o atual apresenta o menor erro absoluto
ciclosimilar: Pesos de cada mês de predição do ciclo do bairro mais similar à parte de referência de atual, entre todos os bairros existentes

Após rodar a função observamos que foram criadas 6 novas colunas:

nrow(casos): [1] 26880
head(casos03[casos03$bairro == "Copacabana" & casos03$caso_anomes == 200201,],1):

ciclosimilar	totultmesref	desviopadraoref	fatorpred_ultref	fatorpred_mesant	ciclomenorerro
631.5	23	10.62085	4.346676	4.346676	148515

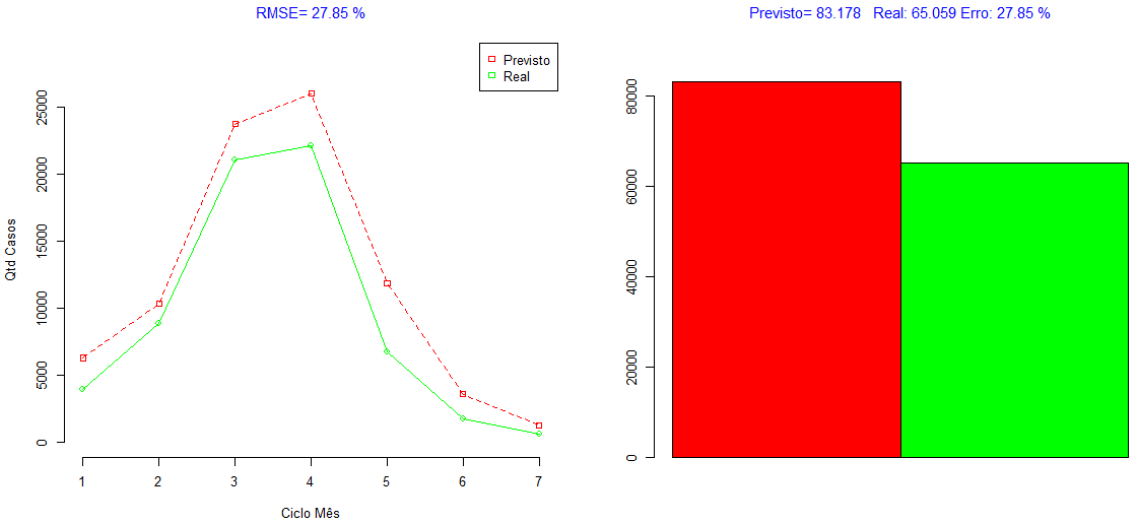
ciclo	bairro	caso_anomes	ciclo_mes	caso_ano	caso_mes	regiao	zona
ciclo2002	Copacabana	200201	6	2002	1	V RA - Copacabana	Zona Sul

area	area_grupo	bairro_ha	bairro_pop2010	ciclo_num	ciclo_parte	qtd_casos
AP 2.1	AG 2	410.09	146392	2002	predicao	452

Processo 4 – Teste final:

```
train04 <- subset(casos03, ciclo_num != ciclo_alvo | (ciclo_num == ciclo_alvo & ciclo_mes <= ciclo_mes_prev))
train04 <- subset(train04, ciclo_num != ciclo_teste | (ciclo_num == ciclo_teste & ciclo_mes <= ciclo_mes_prev))
train04 <- subset(train04, ciclo_parte == "predicao")
train04 <- train04[order(train04$qtd_casos,train04$bairro,train04$caso_anomes),]
test04  <- subset(casos03, ciclo_num == ciclo_teste & ciclo_mes > ciclo_mes_prev)
modelopred04 <- criamodelo(train04,"qtd_casos")
predicao04 <- criapredicao(modelopred04,test04)
```

Repetimos o teste, usando agora o **casos03**, que contém as novas *features*, e o resultado o seguinte: [fig03]

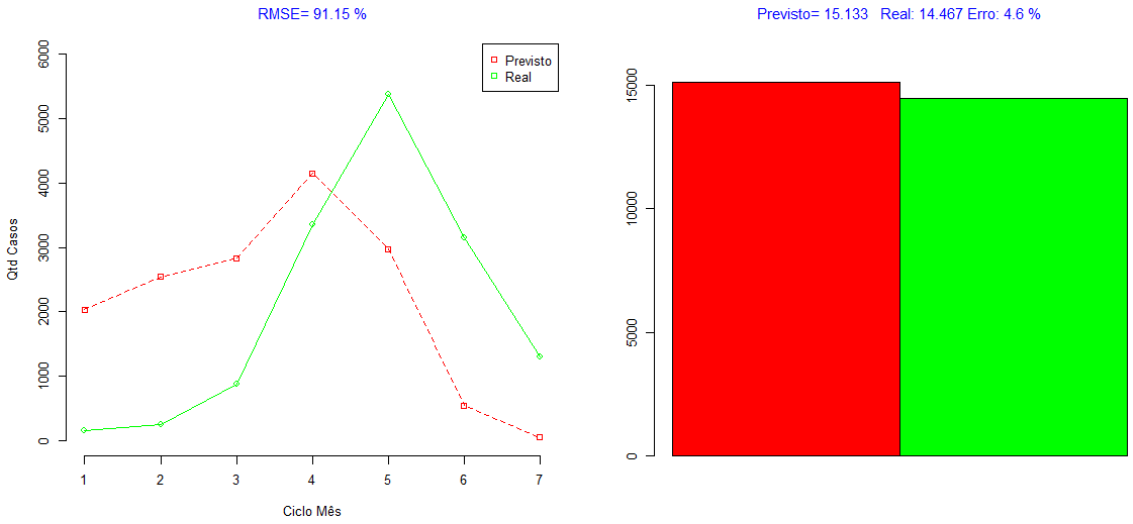


O modelo obtido com as novas *features* demonstrou uma boa aproximação. Apesar de termos conseguido modelos com um RMSE menor, optamos por este porque mostra uma boa aproximação da curva de casos durante os meses.

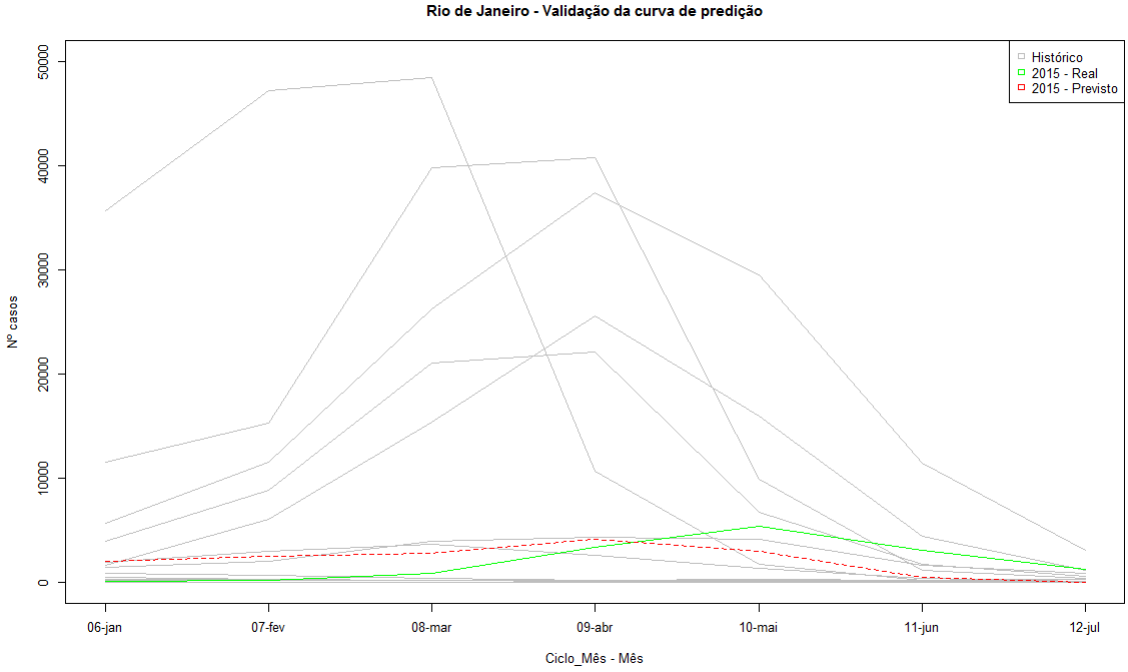
Processo 5 – Validação:

```
valid05  <- subset(casos03, ciclo_num == ciclo_alvo & ciclo_mes > ciclo_mes_prev)
predicao05 <- criapredicao(modelopred04,valid05)
```

Agora que temos nosso modelo de predição, resta testar com os dados reais de 2015, ao que obtivemos o seguinte resultado final:



Cabe ainda comparar nossa predição com as demais curvas históricas:



É possível concluir que conseguimos chegar a um resultado satisfatório de modelo de predição utilizando poucos dados como fonte. Para um total de 14.467 casos reais ocorridos no ciclo de predição de 2015, o modelo criado previu 15.133 casos, com um erro de 4,6 %. A curva de distribuição entre os meses e o pico de casos também apresentou boa similaridade, considerando que os dados usados como fonte eram bastante limitados.

Conclusão:
Demonstramos como um algoritmo de *Machine Learning* tem a capacidade de fazer um modelo de predição a partir de um conjunto de dados puro, sem a necessidade de configurações ou estruturações longas e complexas das variáveis independentes. Ao chamar a função de criação do modelo de predição, apenas passamos para o algoritmo o conjunto de dados e qual das colunas queremos prever. Ao solicitar a previsão, apenas passamos o modelo criado e um conjunto de dados com as mesmas colunas que o usado para a previsão, preenchido com os valores correspondentes ao que desejamos prever.

É certo que trata-se apenas de um exercício de demonstração, uma vez que se encontram disponíveis muitos outros fatores causais que influenciam no ciclo de dengue, mas é de se esperar que implementações mais completas destes métodos entreguem um poder de predição e simulação muito considerável, a um baixo custo de desenvolvimento.

8 – Proposta de solução

Ressaltamos então que o cenário atual de crescimento da disponibilização de dados em saúde no Brasil somado à demanda urgente de ações efetivas de entendimento e apoio logístico ao controle de epidemia de dengue demandam fortemente para a complementação de sua análise os recursos que as tecnologias de *Big Data* nos oferecem.

Isto poderá se dar na montagem de equipes e/ou salas de situação e controle com uso de *Big Data*, contando com cientistas de dados capacitados e foco na criação de um ambiente adequado para o desenvolvimento destas ferramentas.

Estas equipes podem ter uma atuação delimitada em municípios de médio/alto porte, municípios de alto grau de risco, grupos de municípios em áreas de infestação, regiões metropolitanas ou estados da federação, suportando as ações de campo e atendimento com informações preciosas.

Estes esforços de análise poderão contar com as fontes de dados já disponíveis criando um ambiente adequado para a centralização e integração das informações e potencializando o resultado que elas já proporcionam:

- LIRAa - Levantamento Rápido do Índice de Infestação por *Aedes aegypti*
- Plataforma DATASUS
- Dados municipais cadastrais / Cadastro IPTU
- Ações de prevenção e combate a dengue realizadas
- Dados meteorológicos observacionais e previsões
- Pesquisas de campo relacionadas ao tema
- Informações de atendimento e relacionamento com cidadão
- Rastreamento de mídias e redes sociais

A unificação adequada das informações destas várias fontes sob a ótica do *Big Data* pode proporcionar várias aplicações entre as quais destacamos:

- Centralização das informações relativas ao tema e publicação padronizada de dados
- Mecanismo de consulta detalhada do peso dos fatores causais e graus de risco no tempo e espaço
- Simuladores e predições de maior precisão
- Fonte de auxílio à logística das ações e operações
- Fonte de conteúdo assertivo para as ações de mobilização e relacionamento com o cidadão
- Apoio na tomada de decisões dos gestores dos processos

9 – Anexos

Os arquivos citados neste trabalho encontram-se disponíveis para download através do link: https://github.com/ericomartins/bigdatasaude.git

Anexo 01 – csv com dados brutos utilizados
Anexo 02 – código em R da elaboração dos gráficos utilizados
Anexo 03 – código em R dos processos citados no trabalho
Anexo 04 – código em R com as funções chamadas pelo código do anexo 03.

10 – Referências

1. Rita de Cássia CL, Ricardo CGS - Percepção dos usuários sobre o processo de acesso a dados sobre a saúde em sítios do governo federal - XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2013)

2. Chiavegatto Filho ADP - Uso de big data em saúde no Brasil: perspectivas para um futuro próximo - Epidemiol. Serv. Saúde, Brasília, 24(2): 325-332, abr-jun 2015

3. http://www.theguardian.com [Internet] - Good data can help diagnose the health of cities around the world.
   Disponível em: http://www.theguardian.com/healthcare-network/2015/oct/12/good-data-health-cities

4. Secretaria de Vigilância em Saúde − Ministério da Saúde - Boletim Epidemiológico - Volume 46 - N° 44 – 2015

5. Presidência da República - Blog do Planalto [Internet] - Presidenta Dilma anuncia Plano Nacional de Enfrentamento à Microcefalia.
   Disponível em:
   http://blog.planalto.gov.br/presidenta-dilma-anuncia-plano-nacional-de-enfrentamento-a-microcefalia/

6. Vivaldo José BRETERNITZ, Leandro Augusto SILVA − Big Data: um novo conceito gerando oportunidades e desafios - Revista RETC – Edição 13ª, outubro de 2013, página 106

7. Prefeitura do Rio de Janeiro – RJ [Internet] - Casos de Dengue por bairro e período
   Disponível em: http://www.rio.rj.gov.br/web/sms/dengue-casos-bairro-periodo/

8. Tong He - [Internet] - XGBoost- eXtreme Gradient Boosting
   Disponível em: https://github.com/dmlc/xgboost

9. Wikipedia [Internet] - Root-mean-square deviation
   Disponível em: https://en.wikipedia.org/wiki/Root-mean-square_deviation