Forecast de recursos computacionais

Autores: Ana Luiza Chagas de Freitas, Fernando Marques, Gabriel R. Saraiva, Gabriel Sanchez, Henrique Piassi Soares da Silva, Kevin Dias

Introdução

O *forecast* é um método que possibilita o acompanhamento e a previsão de resultados futuros. Considerando o nosso contexto, faremos uma previsão do uso de recursos computacionais de um servidor. Essa previsão é importante pois permite determinar com uma certa antecedência qual será o uso de recursos nesta máquina, identificar possíveis gargalos e embasar a tomada de decisões baseada em dados.

Análise exploratória

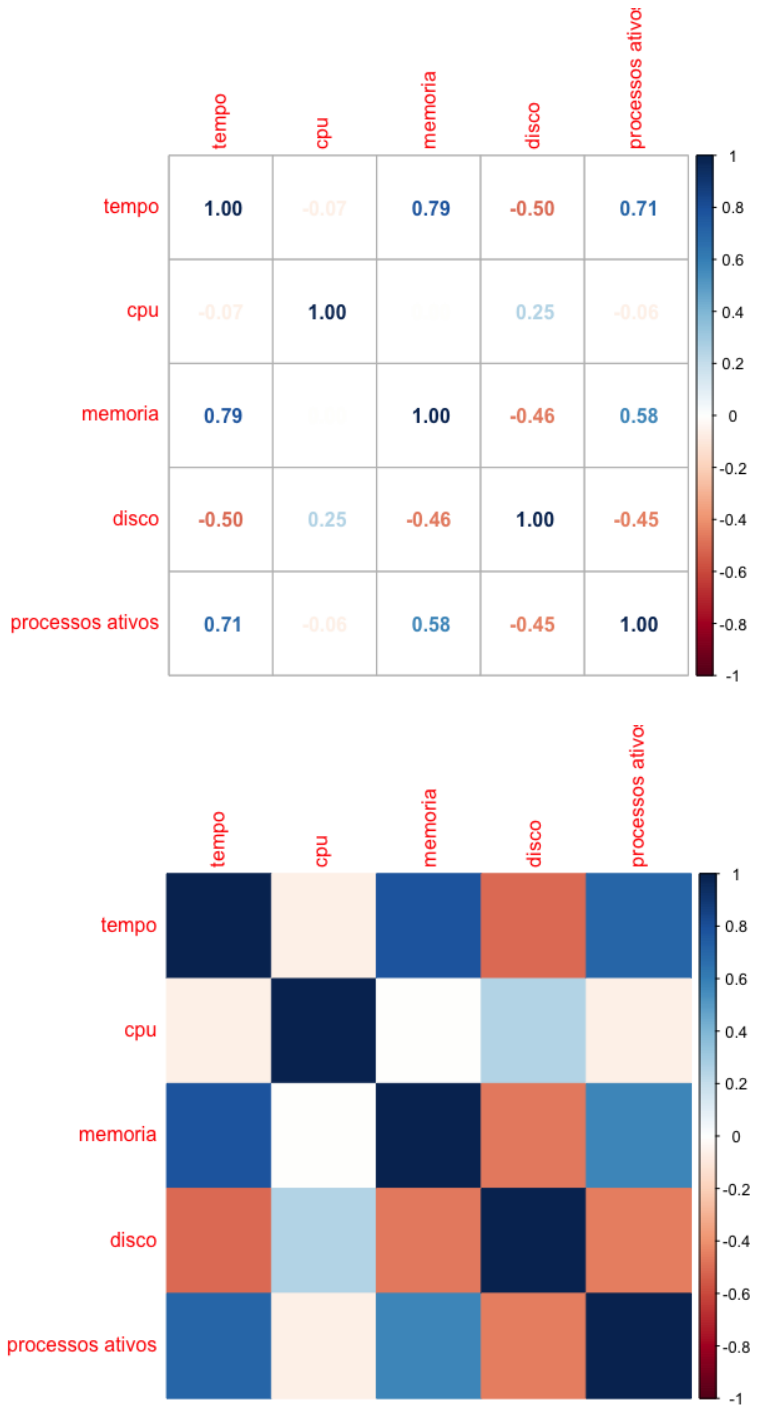
Para a realização deste *forecast*, foram coletados 2451 dados. Os dados coletados foram: data e hora, percentual de uso de CPU, percentual de uso de memória RAM, percentual de uso de disco, processos ativos na máquina e o usuário atual. Foi estabelecido um intervalo padrão de 30 segundos entre cada coleta. Sendo assim, o tempo total para coleta dos dados foi de 20 horas e 26 minutos (das 11:52 do dia 20/11/2021 até as 08:18 do dia 21/11/2021).

	id	dataHora	cpu	memoria	disco	processosAtivos	usuarioAtual	fkServidor
1	1	2021-11-20 11:52:56	13.3	59.4	55.8	412	anafreitas	1
2	2	2021-11-20 11:53:26	3.7	59.4	55.8	407	anafreitas	1
3	3	2021-11-20 11:53:56	4.0	59.5	55.8	406	anafreitas	1
4	4	2021-11-20 11:54:26	4.3	59.3	55.8	409	anafreitas	1
5	5	2021-11-20 11:54:56	4.2	59.2	55.8	411	anafreitas	1
6	6	2021-11-20 11:55:26	1.8	59.4	55.8	406	anafreitas	1
7	7	2021-11-20 11:55:56	1.3	59.4	55.8	406	anafreitas	1
8	8	2021-11-20 11:56:26	4.0	59.3	55.8	406	anafreitas	1
9	9	2021-11-20 11:56:56	34.9	60.0	55.9	417	anafreitas	1
10	10	2021-11-20 11:57:26	46.0	60.0	56.2	411	anafreitas	1
11	11	2021-11-20 11:57:56	31.7	59.4	55.9	426	anafreitas	1

Showing 1 to 11 of 2,451 entries, 8 total columns

Ressalto que a data e hora da coleta foram convertidos para *unix timestamp* nas próximas etapas, a fim de facilitar o uso dessa variável para comparações e cálculos matemáticos. Essa técnica é amplamente utilizada no mundo da computação.

Após esse processo, levantamos a hipótese de que existe uma correlação linear entre as variáveis nos dados coletados. Por exemplo, acreditamos ser possível haver uma correlação entre o horário da coleta e o percentual de uso de memória. Para comprovar essa teoria, utilizamos o Coeficiente de Correlação de *Pearson* para identificar correlação linear entre as variáveis, obtendo o gráfico a seguir:

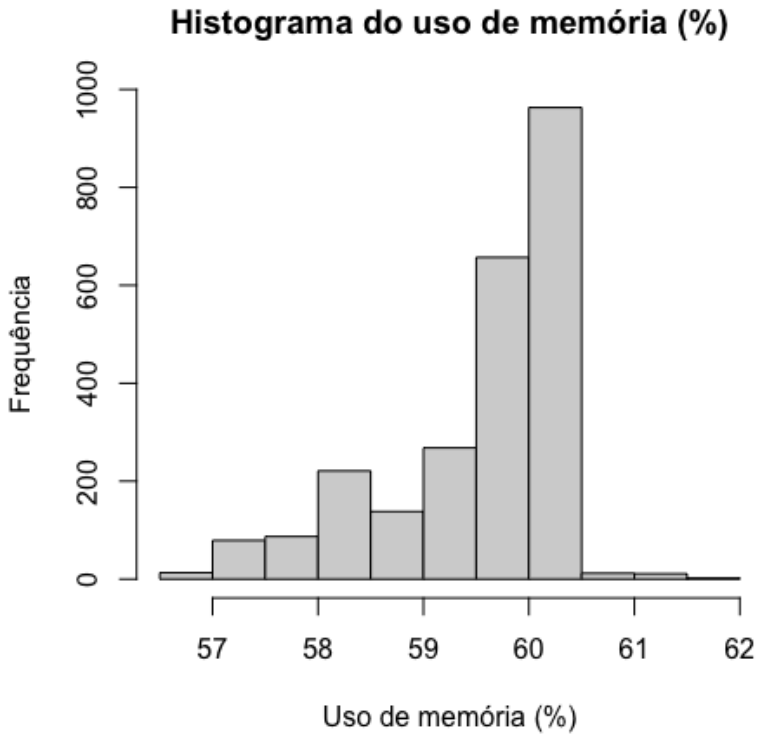


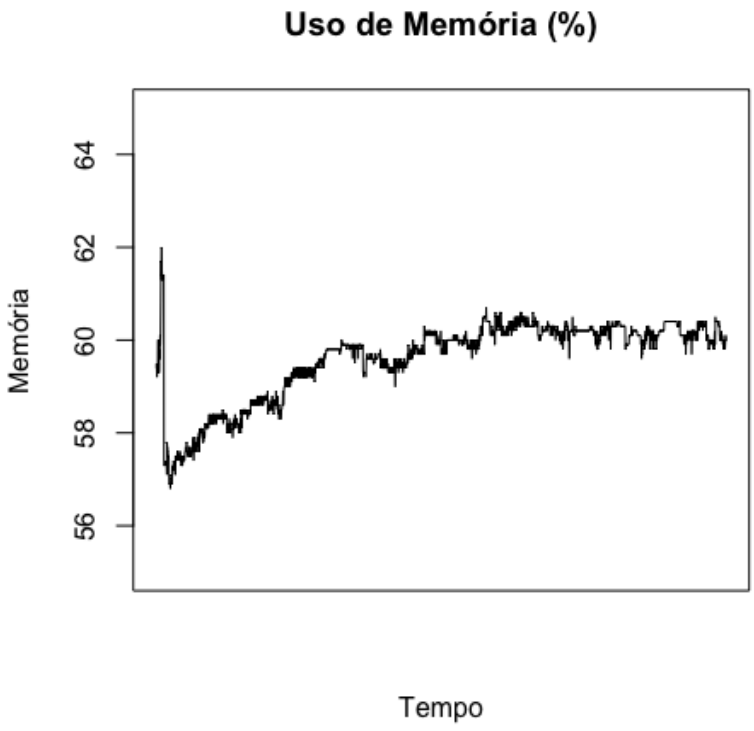
Após análise da correlação, determinamos que as variáveis a serem utilizadas para este *forecast* seriam: memoria x tempo (correlação de 0.79) e processos ativos x tempo (correlação de 0.71).

Em seguida, usamos as variáveis que tiveram alta correlação entre si para fazer um estudo mais aprofundado sobre cada uma delas. Ou seja do uso de memória e dos processos ativos ao longo do tempo.

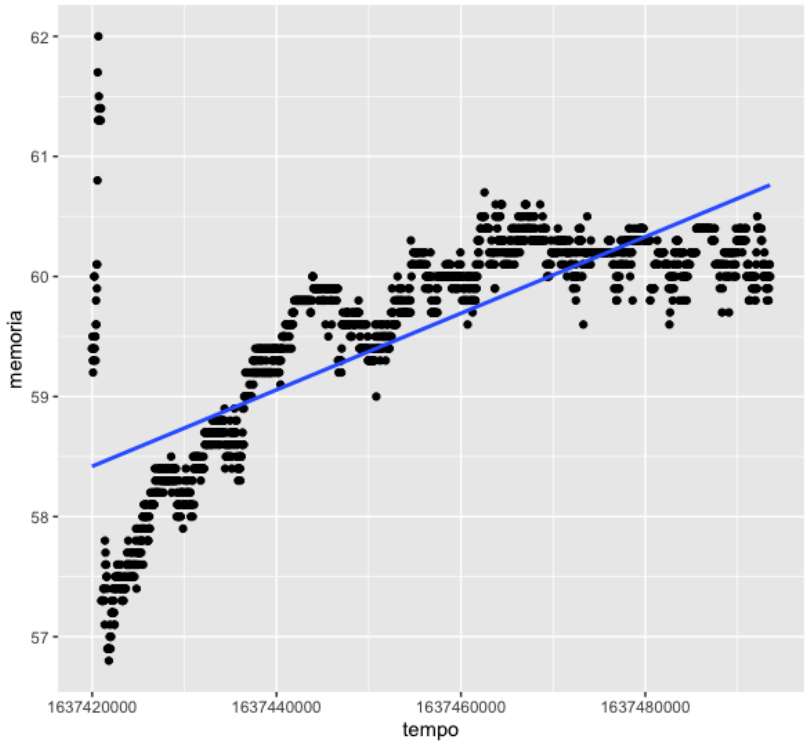
- Memória:

Nessa etapa, plotamos o histograma e o gráfico de linhas para verificar a distribuição por frequência dessa varável e facilitar a visualização do comportamento dos dados de uso de memória ao longo do tempo.



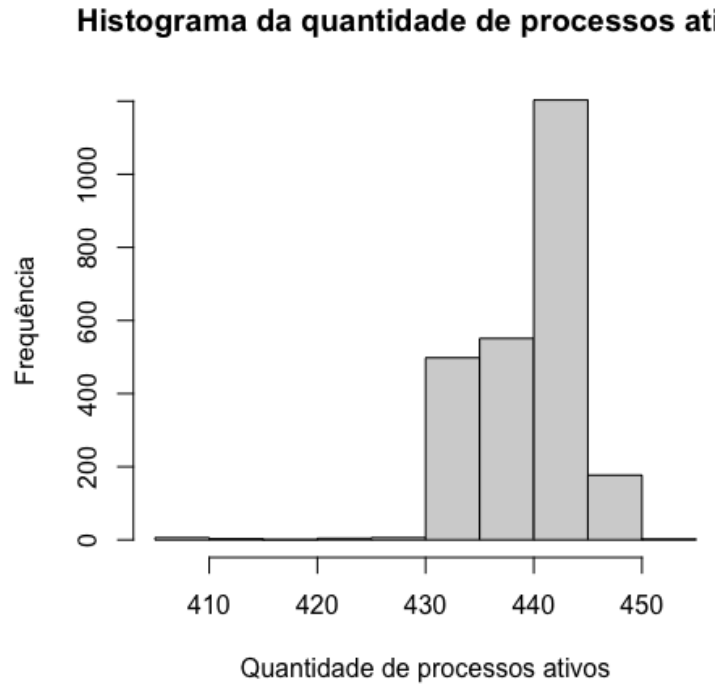


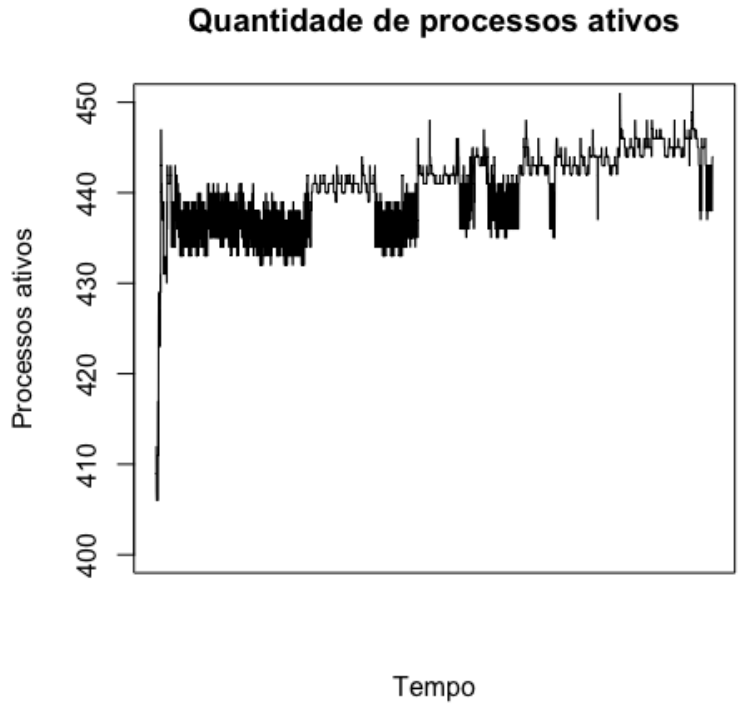
E, por fim, plotamos este último gráfico para ajudar a explicar este conjunto de dados através de uma distribuição linear.



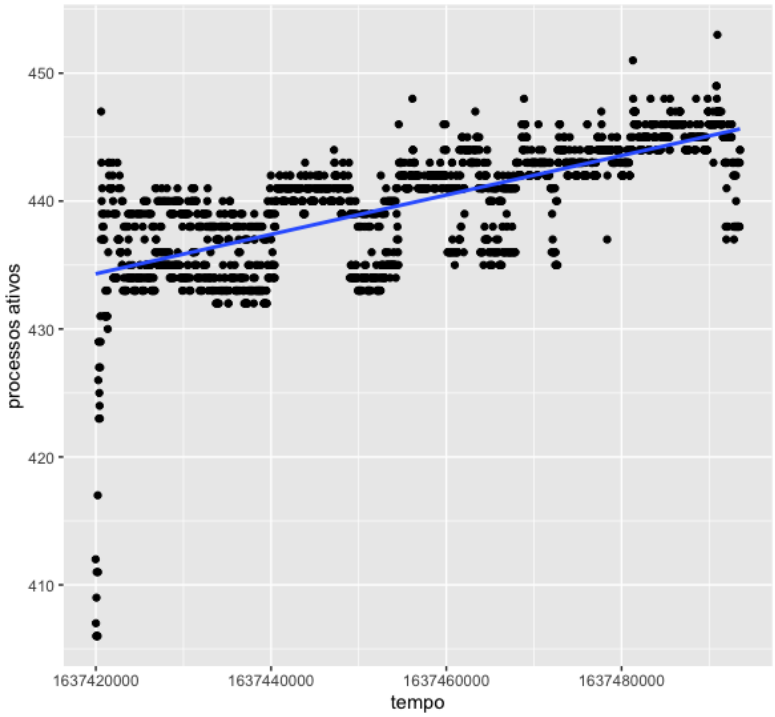
- Processos ativos:

Agora em relação a quantidade de processos ativos, plotamos o histograma e o gráfico de linhas para verificar a distribuição por frequência dessa varável e para facilitar a visualização do comportamento destes dados longo do tempo.





E, por fim, plotamos este último gráfico para ajudar a explicar este conjunto de dados através de uma distribuição linear.



O próximo passo foi determinar qual equação matemática representa cada um dos nossos modelos

- Para memória x tempo:

```
> modelo_memoria <- lm(data$memoria~data$timestamp)
> modelo_memoria

Call:
lm(formula = data$memoria ~ data$timestamp)

Coefficients:
   (Intercept)  data$timestamp
    -5.211e+04       3.186e-05
```

A equação ficou assim:

$$\boldsymbol{memoria} = \mathbf{0.03186}(\boldsymbol{timestamp}) - \mathbf{52110000}$$

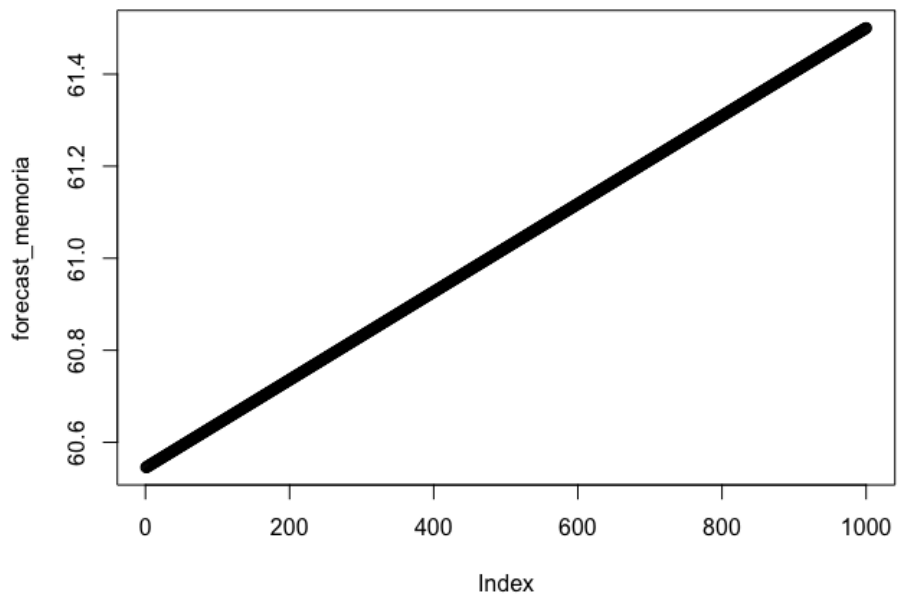
Agora podemos fazer a previsão para as próximas horas. Para isso, foi criada a seguinte função, a fim de trazer uma automação e facilitar previsões futuras:

```
# FORECAST MEMÓRIA

previsao_memoria = function(timestamp){
  valor = 0.03186*timestamp - 52110000
  return(valor) * (10**-3)
}
```

Fazendo a previsão das próximas 8h de uso da memória obtivemos o seguinte resultado:

```
> previsao_memoria = function(timestamp){
+     valor = (0.03186*timestamp - 52110000) * ((10**-3))
+     return(valor)
+ }
> forecast_memoria <- c()
> for(i in 1:1000){
+     add_time = 30*i
+     value=previsao_memoria(1637493554 + add_time)
+     forecast_memoria <- append(forecast_memoria,value)
+     #print(value)
+ }
> previsao_memoria(1637493554)
[1] 60.54463
> previsao_memoria(1637522354)
[1] 61.4622
```



- Para processos ativos x tempo:

```
> modelo_processos_ativos <- lm(data$processosAtivos~data$timestamp)
> modelo_processos_ativos

Call:
lm(formula = data$processosAtivos ~ data$timestamp)

Coefficients:
   (Intercept)  data$timestamp
    -2.517e+05       1.540e-04
```

A equação ficou assim:

$$\boldsymbol{processos_{ativos} = 0.000154(timestamp) - 251700}$$

Agora podemos fazer a previsão para as próximas horas. Para isso, foi criada a seguinte função, a fim de trazer uma automação e facilitar previsões futuras:

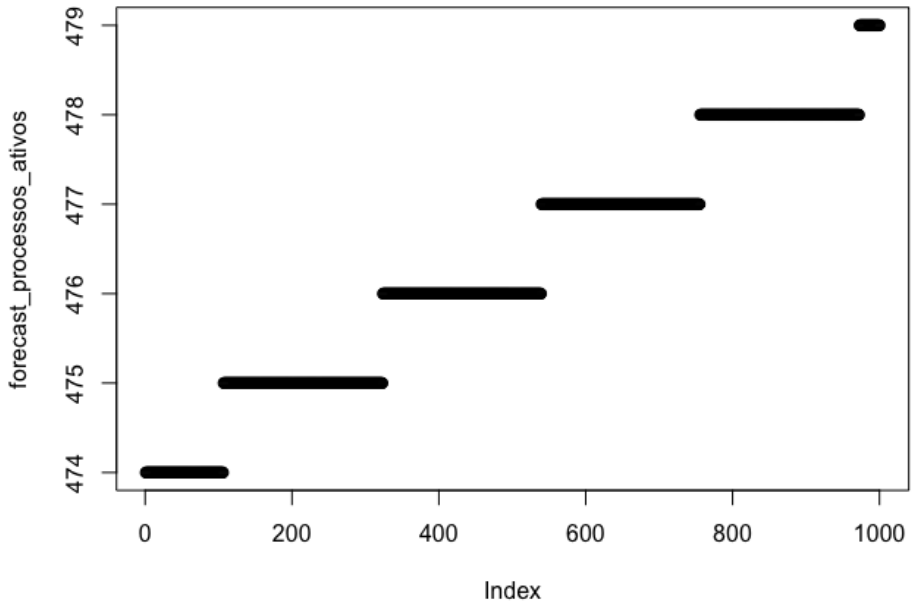
```
# FORECAST PROCESSOS ATIVOS

previsao_processos_ativos = function(timestamp){
  valor = round(0.000154*timestamp - 251700)
  return(valor)
}
```

Observação: a função *round* foi utilizada a fim de evitar o surgimento de uma quantidade de processos com casas decimais, o que não faria sentindo para essa variável.

Fazendo a previsão das próximas 8h da quantidade de processos ativos obtivemos o seguinte resultado:

```
> previsao_processos_ativos = function(timestamp){
+     valor = round(0.000154*timestamp - 251700)
+     return(valor)
+ }
> forecast_processos_ativos <- c()
> for(i in 1:1000){
+     add_time <- 30*i
+     value=previsao_processos_ativos(1637493554 + add_time)
+     forecast_processos_ativos <- append(forecast_processos_ativos,value)
+     #print(value)
+ }
> previsao_processos_ativos(1637493554)
[1] 474
> previsao_processos_ativos(1637522354)
[1] 478
```



Word Cloud dos Processos Ativos

A fim de verificar qual o nome dos processos ativos e qual sua distribuição por frequência, foi plotada uma nuvem de palavras. Dessa forma, obtém-se uma visualização rápida de qual processo ficou em execução por mais tempo nesta máquina.

Conclusão

Fazer o *forecast* para o nosso problema de negócio é estratégico quando pensamos no monitoramento de recursos e na previsibilidade de situações de alerta. Entretanto, embora tenhamos feito uma primeira versão com uma regressão linear simples, entendemos que próximos passos de desenvolvimento devem estar ligados com modelos mais robustos e que considerem relações não lineares entre as variáveis.