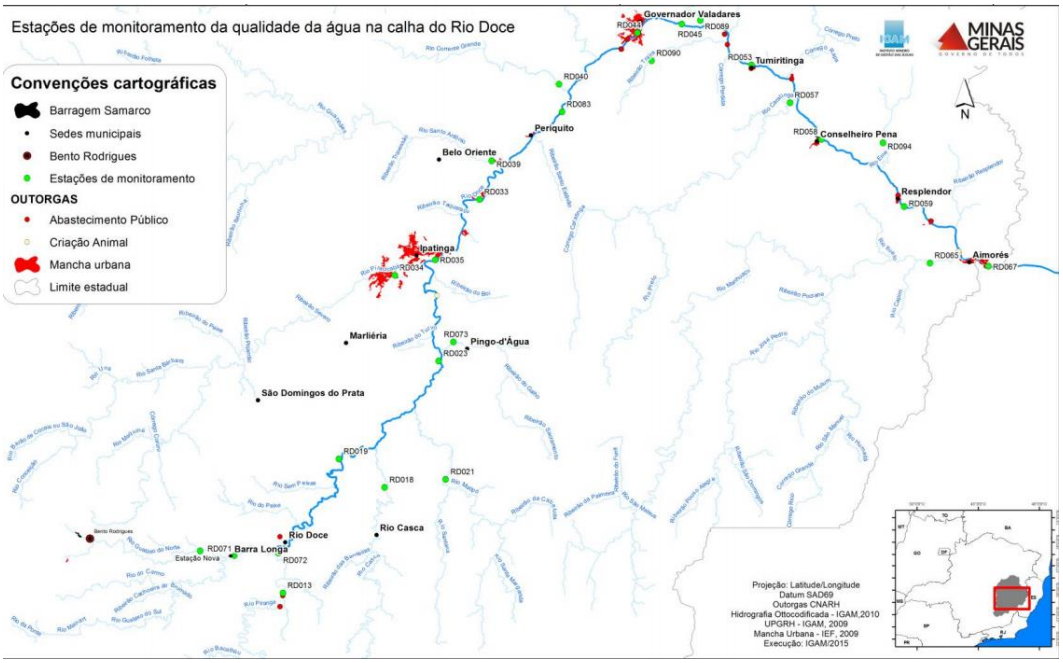
O Dataset escolhido para o trabalho foi o monitoramento da qualidade das águas na bacia hidrográfica do Rio Doce realizado pelo IGAM - Instituto Mineiro de Gestão das Águas, por meio do Programa Águas de Minas que deu início no ano de 1997.

Este monitoramento contempla atualmente 64 estações de amostragem de água, onde são realizadas coletas e análises laboratoriais com periodicidade trimestral e avaliação de aproximadamente 50 parâmetros físico-químicos e hidrobiológicos em 12 estações de monitoramento localizadas na calha do rio Doce, essas coletas e análises são realizadas mensalmente.



O objetivo deste trabalho é verificar se é possível fazer uma comparação pela qualidade da água dos períodos antes, durante e depois do acidente que ocorreu com o Rompimento da Barragem de Mariana em 05 de novembro de 2015, propriedade da SAMARCO, no distrito de Bento Rodrigues.

Da base de dados foi extraído a média do ano de 2014 para gerar os dados antes do acidente, a média de do ano de 2016 para gerar os dados durante o acidente (2014), e a média do ano de 2017 para gerar os dados durante o acidente (2017). Ao longo do trabalho os parâmetros físico-químicos e hidrobiológicos

que foram encontrados no período do acidente não foram encontrados, sendo necessário reduzir o período de análise da base de dados para 6 meses desde o acidente, mas mesmo assim não apresentou resultados satisfatórios sendo necessário então encurtar o período para 3 meses do mês do acidente.

Foi a retirada de 3 pontos de medição do Rio Caratinga, pois já se sabia que as medições destas estações estavam erradas logo no início (Estação RD056 / RD057/ RD093).

Dos 50 parâmetros mencionados, foram selecionados apenas 13, os quais sofreram maior alteração e impactaram mais na qualidade da água, de acordo com a resolução do CONAMA 357 de 2005, tais como: Arsênio (mg/L), Chumbo (mg/L), Ferro (mg/L), Mercúrio (mg/L), Níquel (mg/L), Manganês (mg/L), Coliformes Totais, Demanda Bioquímica de Oxigênio (DBO), Demanda Química de Oxigênio (DQO), Oxigênio(mg/L), Sólidos Totais, Temperatura da água (ºC), Turbidez. Abaixo apresento um resumo da tabela com as classes de qualidade de água.

		Classe			
		1	2	3	4
Arsenio	mg/L	0,01	0,01	0,033	0,033
Chumbo	mg/L	0,01	0,01	0,033	0,033
Coliformes	100 ml	200	1.000	2.500	2.500
DBO	mg/L	3	5	10	10
DQO	cel/mL	20.000	50.000	100.000	100.000
Ferro	mg/L	0,3	0,3	5	5
Manganes	mg/L	0,1	0,1	0,5	0,5
Mercurio	mg/L	0,000	0,000	0,002	0,002
Niquel	mg/L	0,025	0,025	0,025	0,025
Oxigenio	mg/L	6	5	4	2
Solidos	mg/L	500	500	500	500
Temperatura	oC	20	20	20	20
Turbidez	UNT	40	100	100	100

Figura 1 - Resumo da Classificação da Água Doce (CONAMA 357)

O TRABALHO

Resumo dos Dados obtidos pela base de dados

2014

```
> summary(data)
     Cod              Estacao              Rio              Arsenio             Chumbo            Coliformes          DBO        
 Length:61          Length:61          Length:61          Min.   :0.001000   Min.   :0.005000   Min.   :  3394   Min.   :2.000  
 Class :character   Class :character   Class :character   1st Qu.:0.001000   1st Qu.:0.005000   1st Qu.: 14347   1st Qu.:2.000  
 Mode  :character   Mode  :character   Mode  :character   Median :0.001000   Median :0.005000   Median : 29661   Median :2.000  
                                                          Mean   :0.001177   Mean   :0.005275   Mean   : 38109   Mean   :2.110  
                                                          3rd Qu.:0.001000   3rd Qu.:0.005000   3rd Qu.: 53738   3rd Qu.:2.017  
                                                          Max.   :0.007978   Max.   :0.010600   Max.   :132259   Max.   :4.350  
      DQO              Ferro             Manganes          Mercurio          Niquel            Oxigenio         Solidos       
 Min.   : 6.525   Min.   :0.06452   Min.   :0.02055   Min.   :0.2000   Min.   :0.004000   Min.   :4.750   Min.   : 28.75  
 1st Qu.:11.125   1st Qu.:0.11885   1st Qu.:0.04013   1st Qu.:0.2000   1st Qu.:0.004000   1st Qu.:7.633   1st Qu.: 50.00  
 Median :12.317   Median :0.15507   Median :0.06090   Median :0.2000   Median :0.004000   Median :7.825   Median : 66.75  
 Mean   :12.919   Mean   :0.18581   Mean   :0.09642   Mean   :0.2011   Mean   :0.004100   Mean   :7.803   Mean   : 93.60  
 3rd Qu.:13.925   3rd Qu.:0.22865   3rd Qu.:0.10075   3rd Qu.:0.2000   3rd Qu.:0.004000   3rd Qu.:8.050   3rd Qu.:100.00  
 Max.   :39.975   Max.   :0.54632   Max.   :1.10550   Max.   :0.2670   Max.   :0.006175   Max.   :8.525   Max.   :936.00  
  Temperatura       Turbidez      
 Min.   :20.32   Min.   :  4.39  
 1st Qu.:24.32   1st Qu.: 11.67  
 Median :25.94   Median : 21.39  
 Mean   :25.68   Mean   : 36.54  
 3rd Qu.:26.74   3rd Qu.: 31.32  
 Max.   :32.58   Max.   :659.42  
```

Possuem outliers – optou-se por não remover pois é uma característica medir alterações marcantes em cada Estação.

2015 – 2016 (3 meses) não

```
> summary(data)
     Cod              Estacao              Rio              Arsenio             Chumbo           Coliformes          DBO        
 Length:56          Length:56          Length:56          Min.   :0.001000   Min.   :0.005000   Min.   :  6684   Min.   :2.000  
 Class :character   Class :character   Class :character   1st Qu.:0.001000   1st Qu.:0.005000   1st Qu.: 13875   1st Qu.:2.000  
 Mode  :character   Mode  :character   Mode  :character   Median :0.001000   Median :0.005000   Median : 19864   Median :2.000  
                                                          Mean   :0.001746   Mean   :0.013338   Mean   : 35209   Mean   :2.249  
                                                          3rd Qu.:0.001054   3rd Qu.:0.007681   3rd Qu.: 24196   3rd Qu.:2.263  
                                                          Max.   :0.014295   Max.   :0.094000   Max.   :220295   Max.   :4.400  
      DQO              Ferro             Manganes           Mercurio            Niquel            Oxigenio        Solidos       
 Min.   : 7.50   Min.   :0.1204   Min.   :0.02170   Min.   :0.2000   Min.   :0.004000   Min.   :5.600   Min.   :  27.00  
 1st Qu.:14.86   1st Qu.:0.2481   1st Qu.:0.06384   1st Qu.:0.2000   1st Qu.:0.004000   1st Qu.:7.329   1st Qu.:  69.25  
 Median :21.88   Median :0.3470   Median :0.11785   Median :0.2000   Median :0.004000   Median :7.650   Median :  95.50  
 Mean   :25.10   Mean   :0.4121   Mean   :0.67199   Mean   :0.2086   Mean   :0.008787   Mean   :7.559   Mean   : 542.22  
 3rd Qu.:35.62   3rd Qu.:0.4816   3rd Qu.:0.58937   3rd Qu.:0.2000   3rd Qu.:0.004067   3rd Qu.:7.875   3rd Qu.: 643.44  
 Max.   :57.50   Max.   :1.2080   Max.   :6.06350   Max.   :0.4510   Max.   :0.089000   Max.   :8.700   Max.   :8058.00  
  Temperatura       Turbidez      
 Min.   :21.90   Min.   :    5.48  
 1st Qu.:25.18   1st Qu.:   19.73  
 Median :26.50   Median :   55.58  
 Mean   :26.18   Mean   :  863.68  
 3rd Qu.:27.51   3rd Qu.:  724.41  
 Max.   :31.05   Max.   :11362.00  
```

Possuem outliers– optou-se por não remover pois é uma característica medir alterações marcantes em cada Estação.

Nos dados de 2016 foram omitidos alguns NAs com a função NA.OMIT (4 da variável Arsênio, 3 da variável Chumbo, 3 da variável Níquel, 5 da variável Mercúrio).

2017

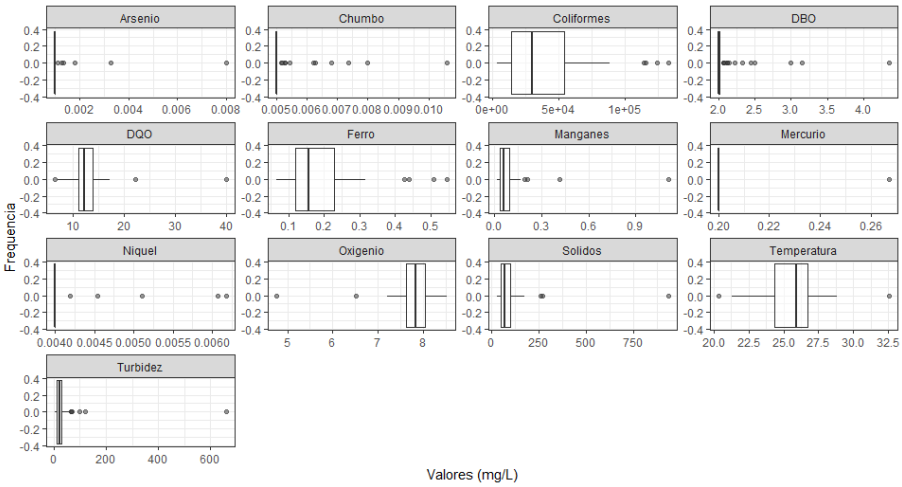
```
> summary(data)
     Cod              Estacao              Rio              Arsenio             Chumbo           Coliformes          DBO        
 Length:62          Length:62          Length:62          Min.   :0.001000   Min.   :0.005000   Min.   : 4227   Min.   :2.000  
 Class :character   Class :character   Class :character   1st Qu.:0.001000   1st Qu.:0.005000   1st Qu.:11484   1st Qu.:2.000  
 Mode  :character   Mode  :character   Mode  :character   Median :0.001000   Median :0.005000   Median :16000   Median :2.000  
                                                          Mean   :0.001516   Mean   :0.005564   Mean   :15520   Mean   :2.194  
                                                          3rd Qu.:0.001042   3rd Qu.:0.005368   3rd Qu.:20817   3rd Qu.:2.000  
                                                          Max.   :0.025800   Max.   :0.012851   Max.   :24196   Max.   :7.933  
      DQO              Ferro             Manganes           Mercurio            Niquel            Oxigenio        Solidos       
 Min.   : 8.333   Min.   :0.1209   Min.   :0.02045   Min.   :0.2000   Min.   :0.004000   Min.   :5.133   Min.   : 46.00  
 1st Qu.:11.400   1st Qu.:0.1932   1st Qu.:0.03703   1st Qu.:0.2000   1st Qu.:0.004000   1st Qu.:7.691   1st Qu.: 66.00  
 Median :13.317   Median :0.2594   Median :0.06678   Median :0.2000   Median :0.004000   Median :7.900   Median : 82.33  
 Mean   :14.719   Mean   :0.2891   Mean   :0.16945   Mean   :0.2016   Mean   :0.004399   Mean   :7.843   Mean   :124.56  
 3rd Qu.:17.731   3rd Qu.:0.3714   3rd Qu.:0.16922   3rd Qu.:0.2000   3rd Qu.:0.004394   3rd Qu.:8.073   3rd Qu.:182.78  
 Max.   :31.000   Max.   :0.8777   Max.   :1.71800   Max.   :0.3015   Max.   :0.007161   Max.   :9.300   Max.   :350.88  
  Temperatura       Turbidez      
 Min.   :19.00   Min.   :  3.63  
 1st Qu.:23.12   1st Qu.: 11.32  
 Median :24.38   Median : 26.81  
 Mean   :24.28   Mean   : 95.23  
 3rd Qu.:25.67   3rd Qu.: 92.02  
 Max.   :27.97   Max.   :692.41  
```

Possuem outliers– optou-se por não remover pois é uma característica medir alterações marcantes em cada Estação.

Boxplot – Mostra a variação de cada variável com seus respectivos outliers

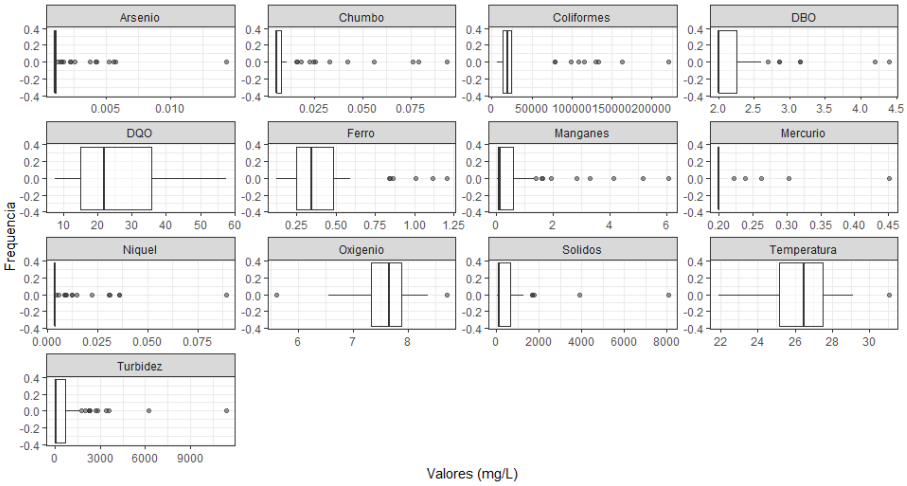
2014

Padrões de normalidade.



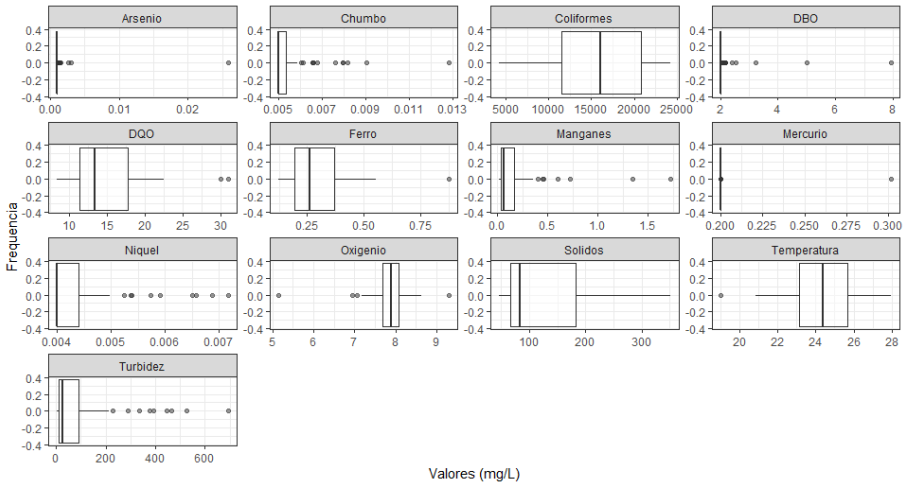
2015 – 2016 (3 meses)

Aumento da Turbidez, Solidos, Manganês, Ferro, Níquel – Bem acima da normalidade.



2017

Retornando os valores para a normalidade, embora a Turbidez, Ferro e Manganês ainda esteja acima do normal.



Matriz de Correlação de das variáveis

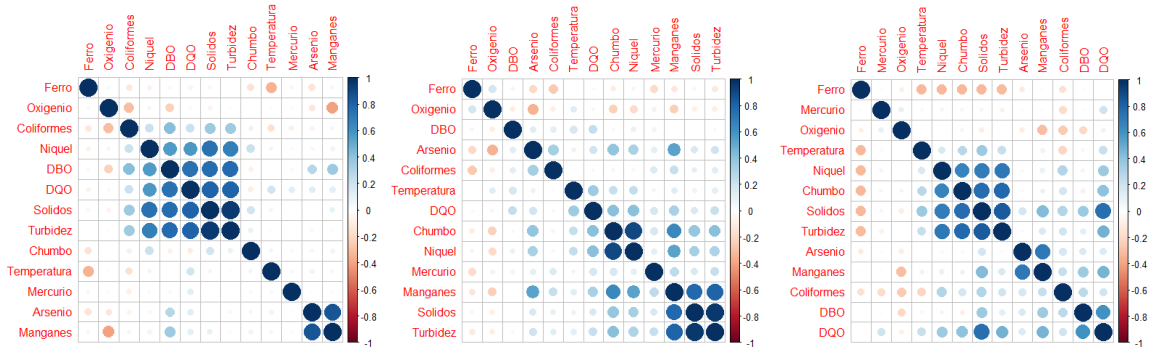
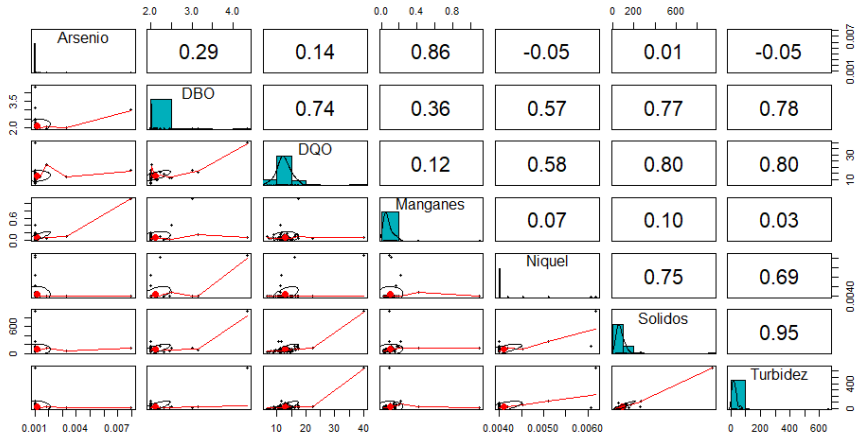


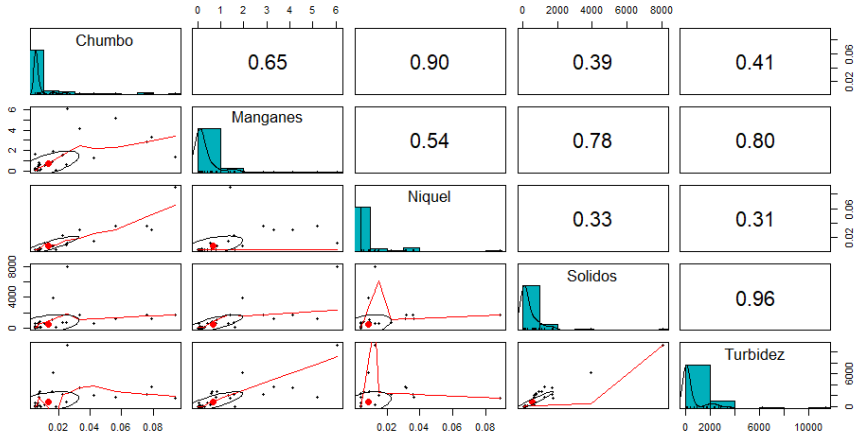
Figura 2 - Matriz de Correlação 2014/ 2015-2016 (3 meses) / 2017

Foi separado as variáveis mais impactantes – valores acima de 0,60.

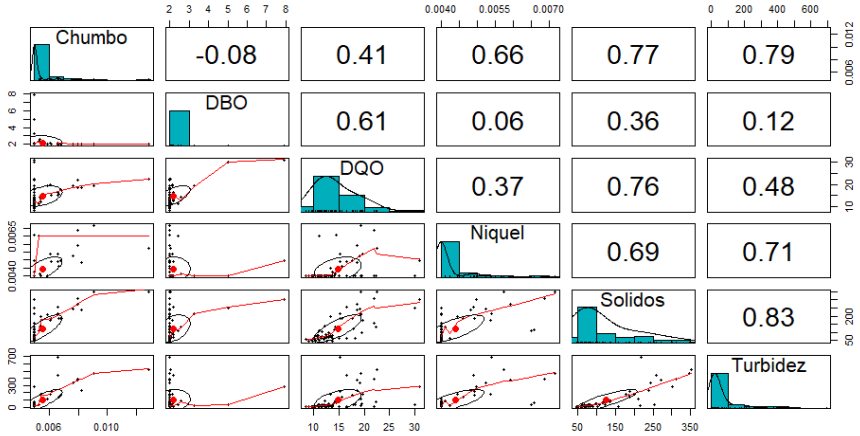
2014



2015 – 2016 (3 meses)



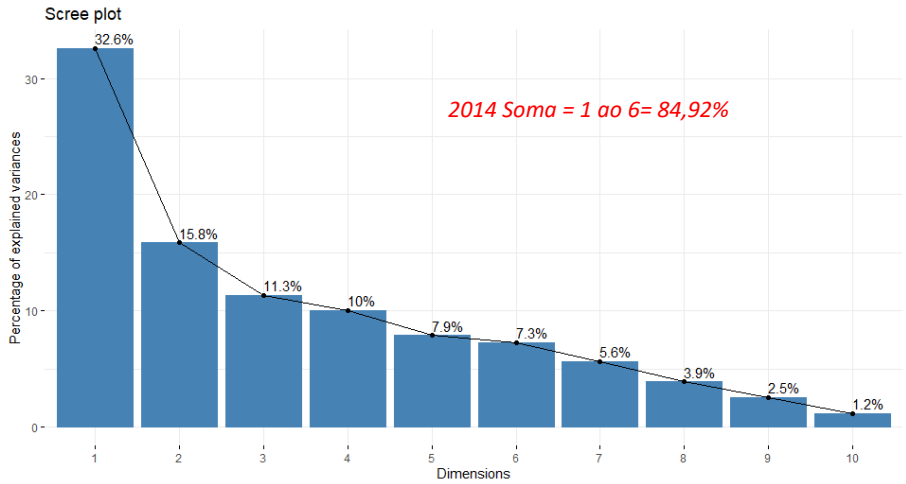
2017

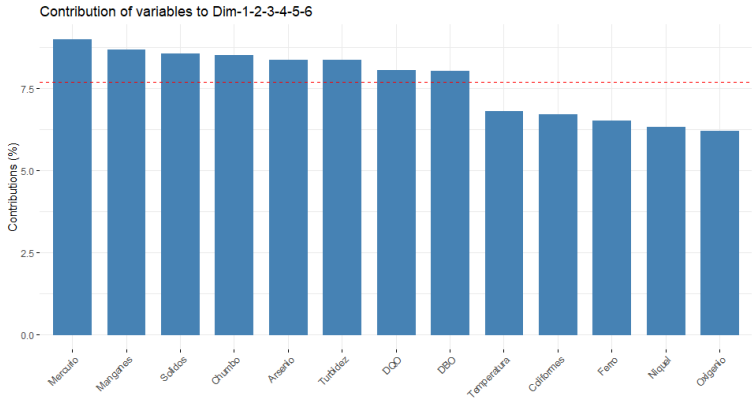


Os dados foram normalizados utilizando PCA para diminuir o ruído (redução de dimensionalidade - variáveis). Nos 3 bancos de dados em questão, a proporção acumulativa no valor de 80% corresponde a 6 variáveis.

2014

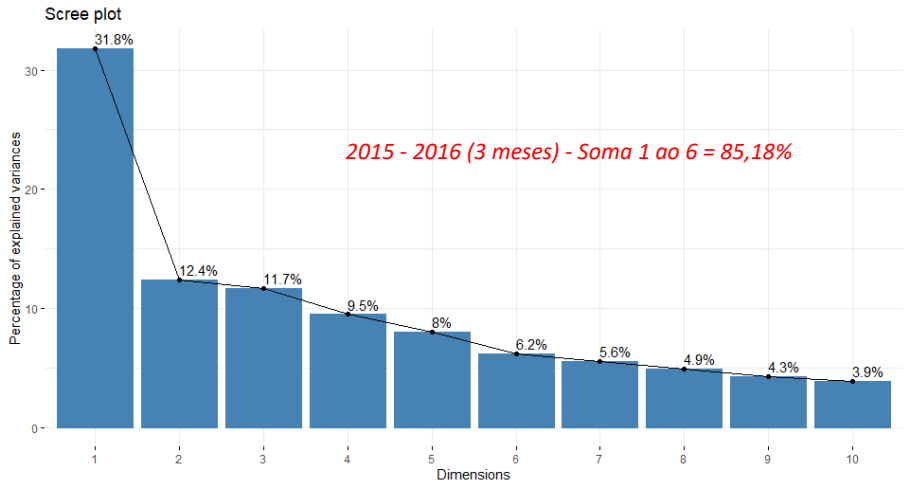
```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     2.0578 1.4353 1.2145 1.1419 1.01099 0.97139
Proportion of Variance 0.3257 0.1585 0.1135 0.1003 0.07862 0.07259
Cumulative Proportion  0.3257 0.4842 0.5977 0.6980 0.77660 0.84919
```

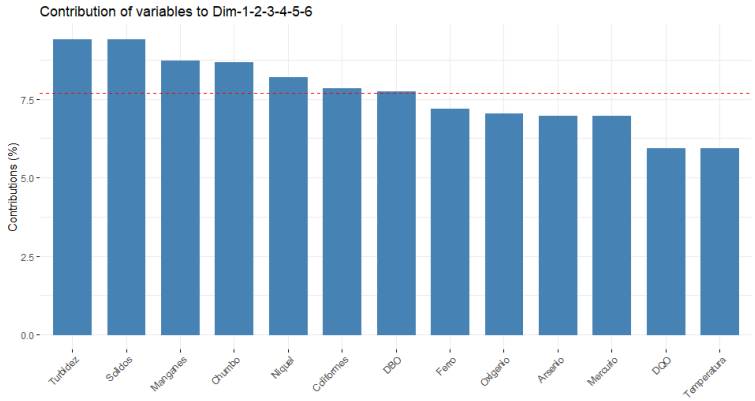




2015 – 2016 (3 meses)

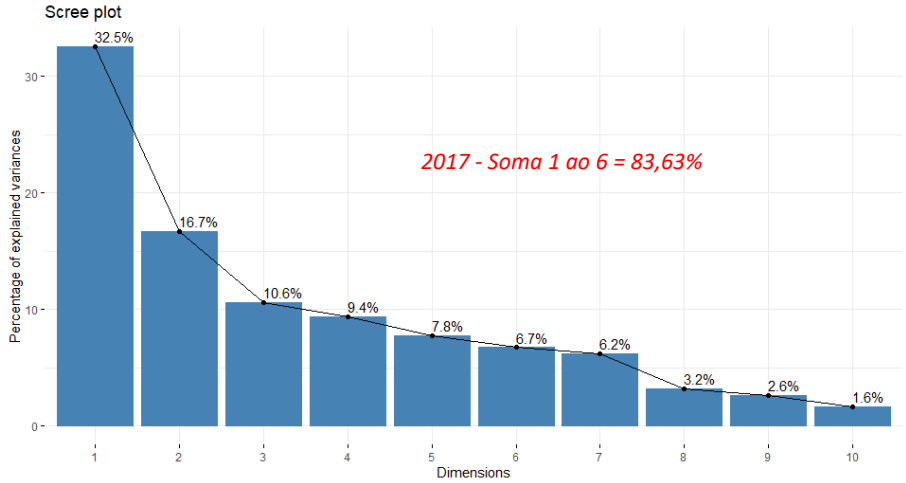
```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.0328 1.2686 1.2343 1.11157 1.02003 0.89789 0.85152
Proportion of Variance 0.3179 0.1238 0.1172 0.09505 0.08004 0.06202 0.05578
Cumulative Proportion  0.3179 0.4417 0.5589 0.65393 0.73396 0.79598 0.85176
```

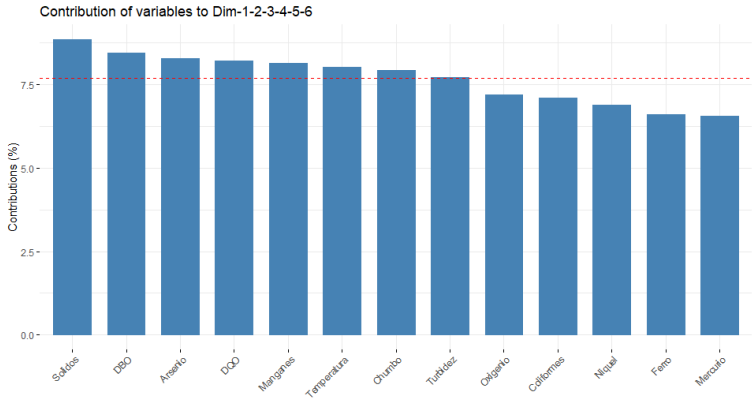




2017

```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     2.0565 1.4718 1.1736 1.10255 1.00458 0.9347
Proportion of Variance 0.3253 0.1666 0.1060 0.09351 0.07763 0.0672
Cumulative Proportion  0.3253 0.4920 0.5979 0.69144 0.76907 0.8363
```





Dos 64 pontos de medição foram retirados 3 do Rio Caratinga, pois foi considerado um outlier logo no ínicio (fora d.

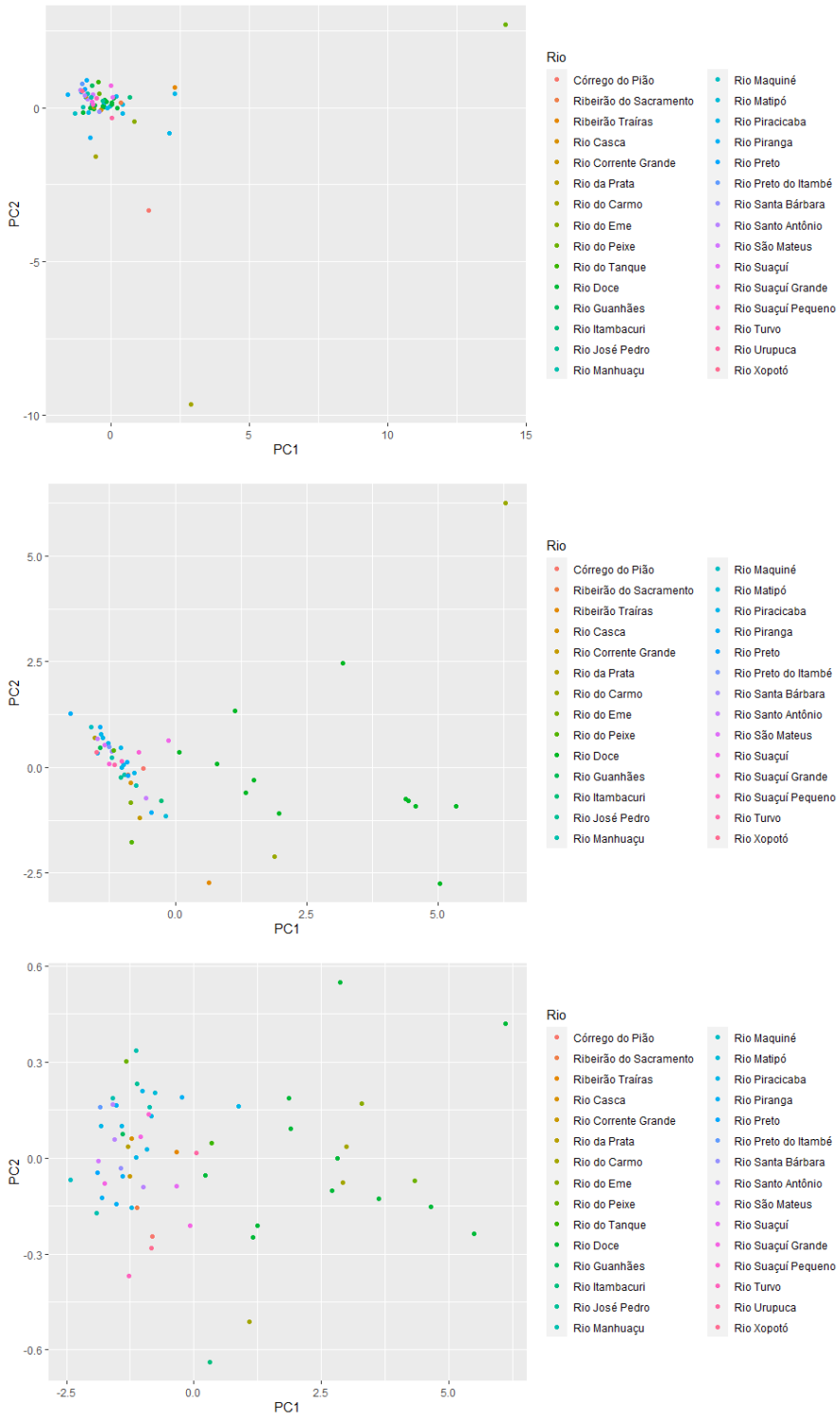


Figura 3 - 2014 / 2016 / 2017 - por tipo de Classe

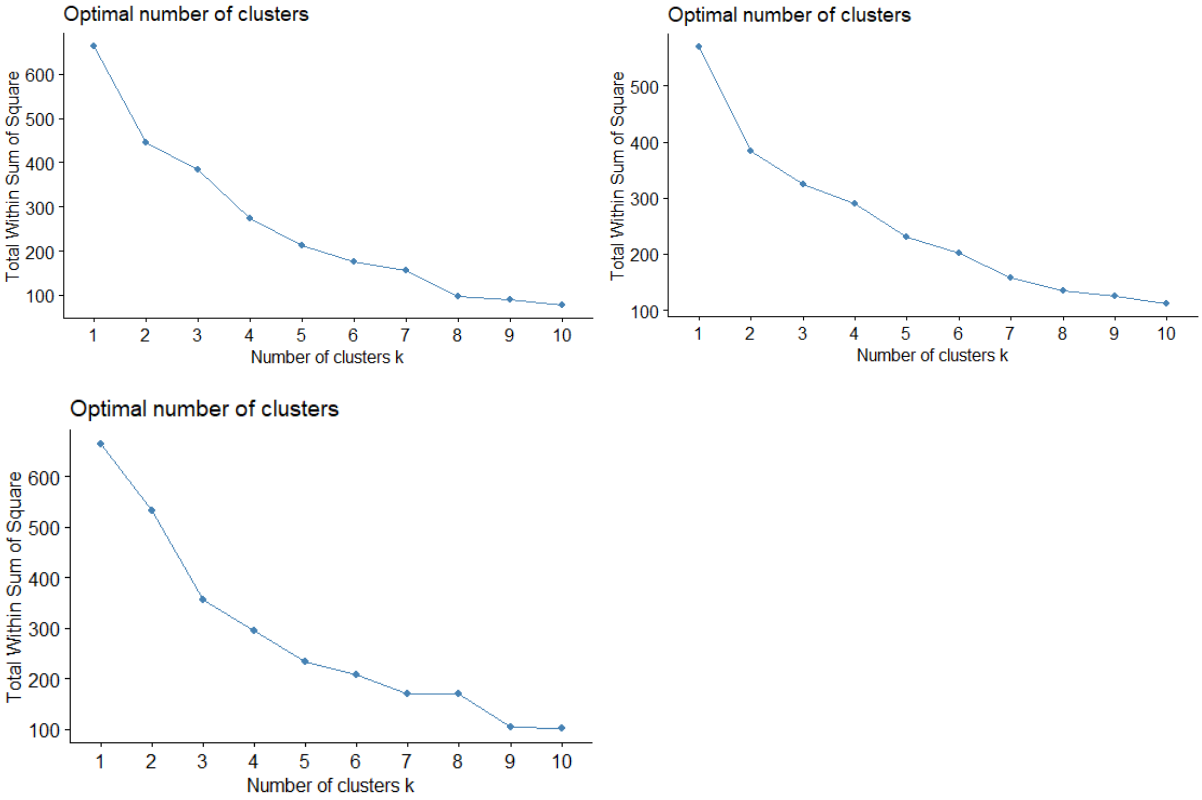


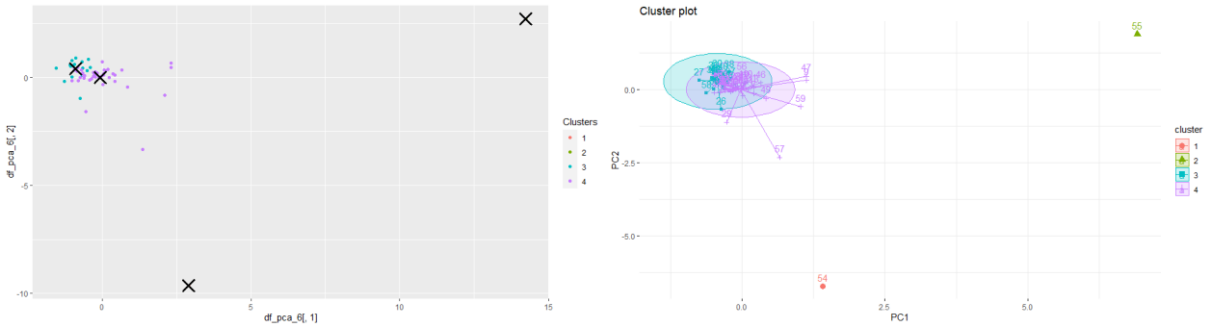
Figura 4 - Número Ideal de Clusters 2014 / 2015-2016 (3 meses) / 2017

Todos indicam um Cluster de 6 pelo método de cotovelo.

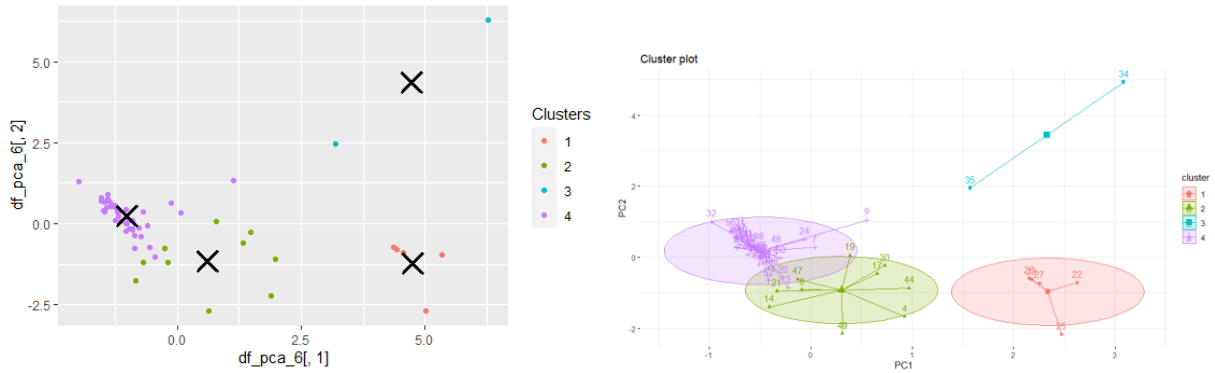
MÉTODOS DE AGRUPAMENTO

1) KMEANS – apesar do método cotovelo sugerir 6 agrupamentos utilizei 4 devido a classificação das águas se enquadrarem em 4 classes.

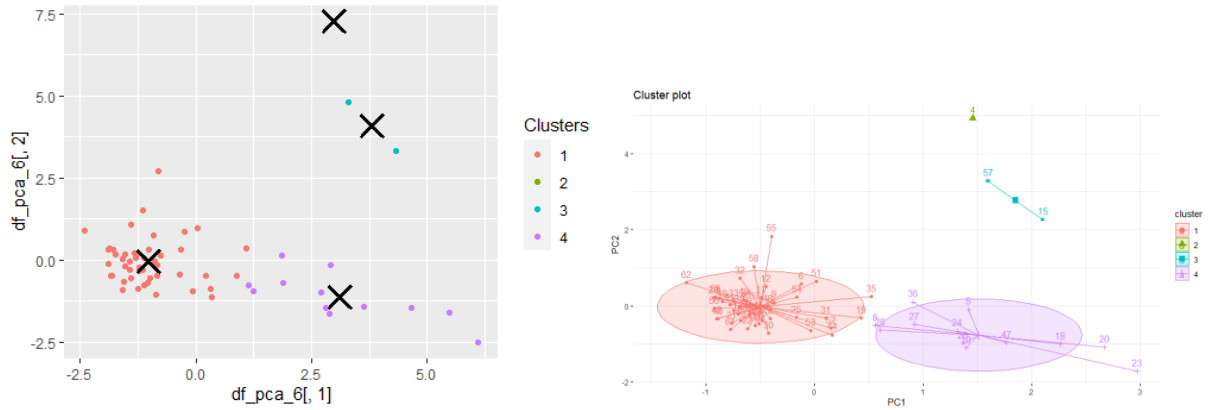
2014 (n=4)



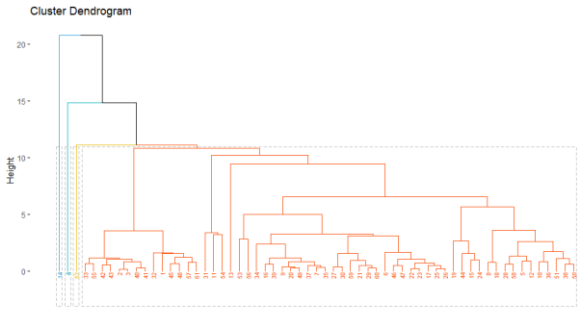
2015 – 2016 (3 meses) (n=4)

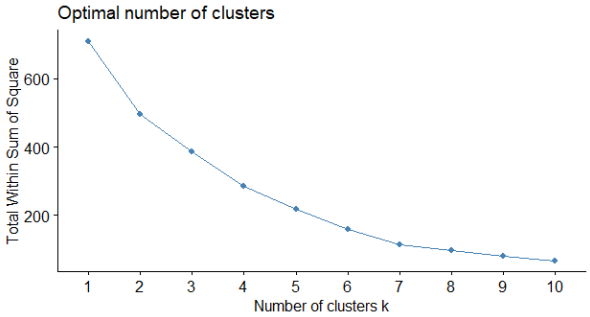


2017 (n=4)

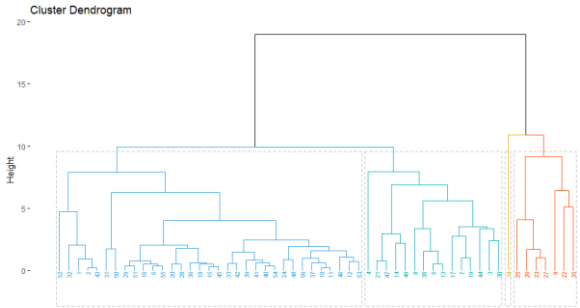


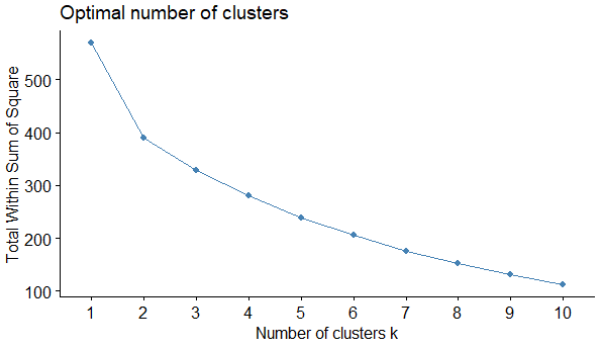
2) HIERÁRQUICO (N = 4)



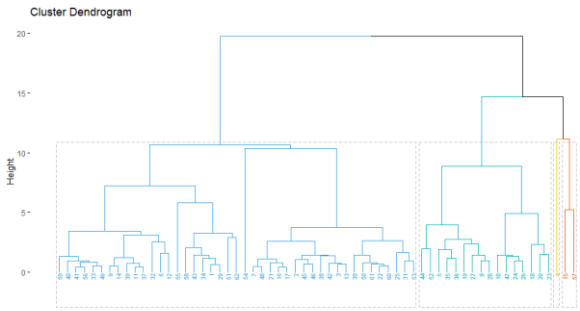


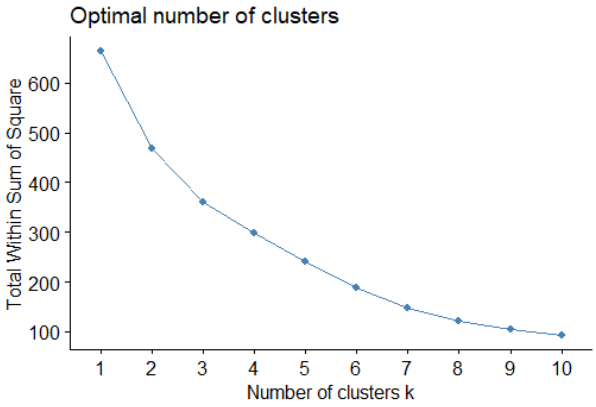
2015 – 2016 (3 meses)



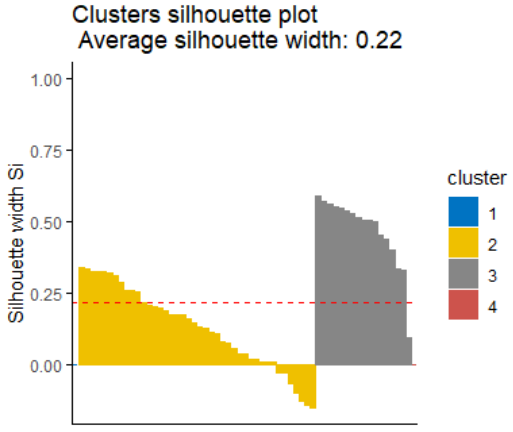


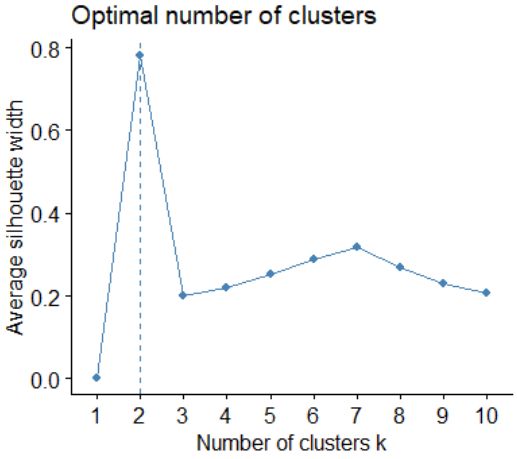
2017





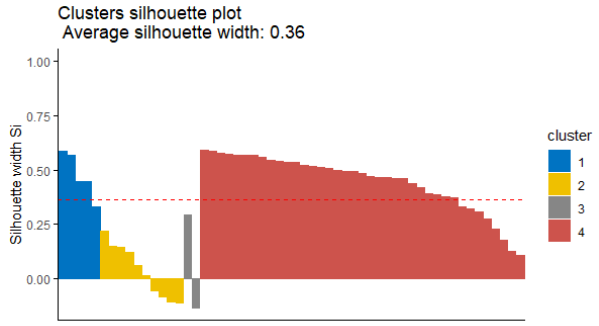
ANÁLISE DA SILHUETA

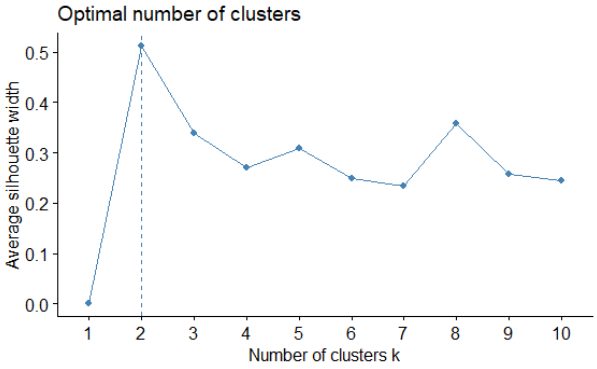





```
  cluster size ave.sil.width
1       1    1          0.00
2       2    1          0.00
3       3    1          0.00
4       4   61          0.77
```

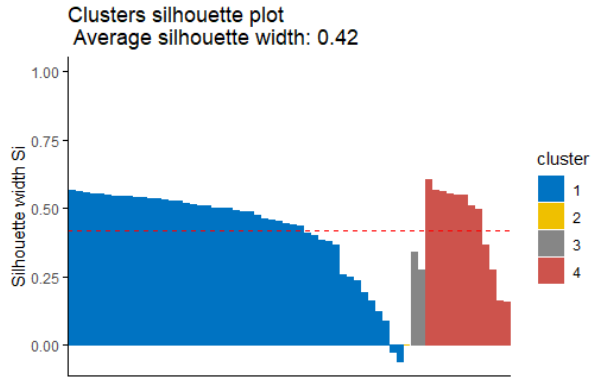
2015 – 2016 (3 meses)

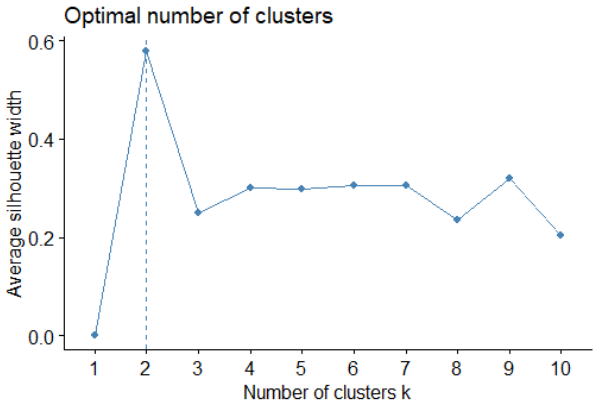





```
  cluster size ave.sil.width
1       1    5          0.47
2       2   10          0.03
3       3    2          0.08
4       4   39          0.44
```

2017






```
  cluster size ave.sil.width
1       1    1          0.00
2       2   42          0.13
3       3   17          0.47
4       4    1          0.00
```

2014

```
> summary(intern)

Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                 2       3       4       5       6

hierarchical Connectivity   2.9290  5.8579  8.7869 11.7159 18.4194
             Dunn           0.9417  0.8366  0.8928  0.5977  0.4108
             Silhouette     0.7796  0.7129  0.6448  0.6022  0.4898
kmeans       Connectivity   2.9290  5.8579  8.7869 14.0615 23.0897
             Dunn           0.9417  0.8366  0.8928  0.2404  0.1591
             Silhouette     0.7796  0.7129  0.6448  0.5093  0.4163
pam          Connectivity   2.9290 22.9885 31.1329 31.5829 33.5496
             Dunn           0.9417  0.0477  0.0405  0.0593  0.0616
             Silhouette     0.7796  0.2014  0.1363  0.1755  0.1909

Optimal Scores:

             Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         0.9417 hierarchical 2
Silhouette   0.7796 hierarchical 2
```

- Vizinhos mais próximos
- Índice do pior caso
- Como melhor se encontra cada objeto em seu conjunto (Mais próximo de 1 – melhor classificado) ideal >0,7

2015 – 2016 (3 meses)

```
> summary(intern)

Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                  2       3       4       5       6

hierarchical Connectivity    2.9290  5.9690 12.7012 16.9913 17.9913
             Dunn            0.6546  0.4444  0.3036  0.3142  0.3142
             Silhouette      0.6181  0.4664  0.4866  0.4990  0.4689
kmeans       Connectivity    9.5234 14.4750 17.1579 19.0869 34.9655
             Dunn            0.2533  0.2785  0.3036  0.3142  0.0866
             Silhouette      0.5290  0.5082  0.4806  0.4893  0.3494
pam          Connectivity   10.4111 26.0980 34.3635 35.5579 37.1560
             Dunn            0.2417  0.0725  0.0725  0.0790  0.0565
             Silhouette      0.5077  0.3548  0.2875  0.3035  0.3139

Optimal Scores:

             Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         0.6546 hierarchical 2
Silhouette   0.6181 hierarchical 2
```

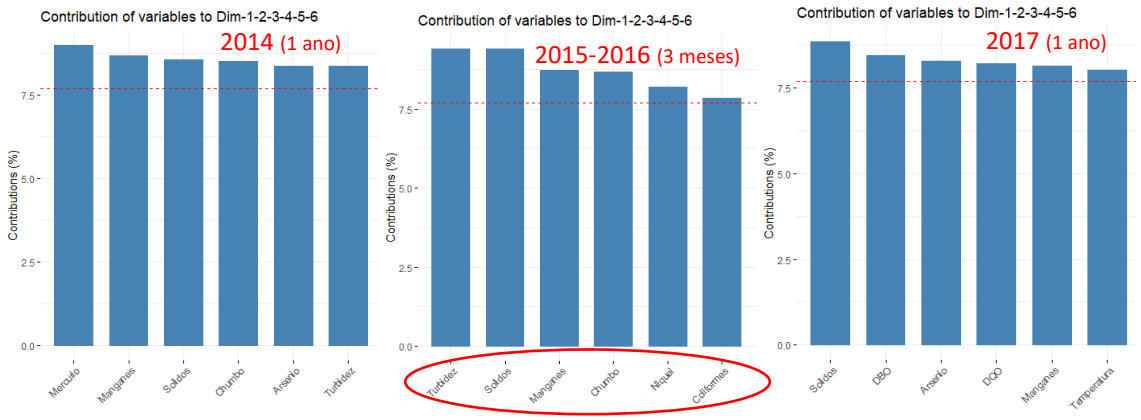
2017

```
> summary(intern)

Clustering Methods:
 hierarchical kmeans pam

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                  2       3       4       5       6

hierarchical Connectivity    2.9290  6.7869  9.7159 15.4750 18.4040
             Dunn            0.7671  0.5362  0.5838  0.2526  0.2757
             Silhouette      0.6195  0.5523  0.5331  0.4813  0.4077
kmeans       Connectivity    5.7869  6.7869  9.7159 18.1179 33.2012
             Dunn            0.4779  0.5362  0.5838  0.1538  0.0823
             Silhouette      0.5779  0.5523  0.5331  0.4333  0.3003
pam          Connectivity    6.5377 18.5171 22.1627 23.1627 25.4583
             Dunn            0.1215  0.0905  0.0942  0.1321  0.1555
             Silhouette      0.4051  0.2504  0.3019  0.3114  0.3179

Optimal Scores:

             Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         0.7671 hierarchical 2
Silhouette   0.6195 hierarchical 2
```

CONCLUSÃO

Devido ao rompimento da barragem os parâmetros que comprovadamente sofreram maior alteração foram: Aumento dos índices de Turbidez, Ferro, Manganês, Níquel e Sólidos em suspensão e Redução dos índices de Oxigênio e DBO. Com relação aos índices de Mercúrio e Arsênio, estes não sofreram alteração, já que os mesmos existiam no sedimento do rio.

Nas bases de dados selecionadas pudemos observar esta mudança de comportamento entre as variáveis. (parâmetros). Sendo observado durante o período do acidente estes parâmetros citados acima como mais importantes a Turbidez, Sólidos, Manganês, Níquel. Confirmando a adequação da base de dados e a algoritmo à realidade.



Conforme mencionado acima, apesar do método cotovelo sugerir agrupamento de 6 foi utilizado agrupamento de 4, observando o critério do CONAMA que classifica em 4 classes a qualidade de águas.

- Antes do rompimento da barragem (2014), o método Kmeans agrupa a maioria das estações em 2 grupos bem próximos, significando estas possuem características semelhantes bem próximas, ficando algumas estações como outliers. Embora o índice silhueta permite criar até 3 grupos, apresentou melhor resultado para 2 grupos, ambos acima de 0,7. Foi sugerido método cluster Hierárquico para 2 grupos ou Kmeans (apresentou o mesmo índice).
- Logo após o acidente (2015 - 2016), o agrupamento Kmeans criou 3 grupos sendo que 2 deles mais próximos e o terceiro um pouco mais distante. Este distante é o que apresentou mais alterações na qualidade de água, referente a algumas Estações do Rio Doce. O índice silhueta apresenta melhor resultado para apenas 2 grupos, porém abaixo de 0,7. Foi sugerido cluster Hierárquico de 2 grupos.
- 1 anos após o acidente (2017), o agrupamento gerado foi em 2 grupos, porém as distâncias entre estes ficaram mais próximas. O índice silhueta não apresentou um resultado satisfatório para 2 grupos. Sugerido cluster Hierárquico de 2 grupos.

Índice abaixo de 0,7 pode indicar que a base de dados não foi bem trabalhada, isto é, possíveis outliers poderiam ter sido retirados para não atrapalhar na classificação, já que o método Kmeans é sensível a outliers. O que foi possível detectar pela presença de clusters com poucos pontos. Porém não foram retirados pois o intuito era analisar todo o trecho da Bacia do Rio Doce.