Análises exploratória dos dados geoquimicos

PJ e Silvia

Última atualização: November 29, 2007

1 Introdução

Este documento trata das análises estatística (inicialmente exploratórias) dos dados de teores de elementos coletados em 698 pontos no estado do Paraná. O objetivo final é compreender a distribuição destes elementos na área para construção posterior de modelos que possam investigar possíveis relações de tais elementos com a ocorrência de neoplasias registradas nos municípios do estado.

As amostras são coletadas em sedimentos de rios (REVISAR/DETALHAR AQUI...) . Desta forma cada dado reflete o teor do elemento na microbacia correspondente. *(PJ: dúvida – todas as microbacias do estado foram amostradas?)* A Figura 1 mostra a região de estudo particionada em microbacias. *(PJ: acrescentar a figura com as bacias e talvez municípios).*

Foi excluído do conjunto de dados originais o dado da posição incidada a seguir, que apresentou teores nulos para todos os elementos, o que sugere que esta é uma amostra com valores não disponíveis. Isto é reforçado pelo fato de que a data da análise não está disponível.

```
    BACIAS_ID LABEL_UTME LABEL_UTMN LONGITUDE LATITUDE  SIGLA
451       472   233226.7    7213220  -53.6468 -25.1734 IG-203
    DATA_ANALI AG_ES2 AL_ES2 B_ES2 BA_ES2 CA_ES2 CD_ES2 CO_ES2 CR_ES2
451       <NA>      0      0     0      0      0      0      0      0
    CU_ES2 FE_ES2 GA_ES2 IN_ES2 K_ES2 LI_ES2 MG_ES2 MN_ES2 MO_ES2
451      0      0      0      0     0      0      0      0      0
    NA_ES2 NI_ES2 PB_ES2 SR_ES2 TL_ES2 V_ES2 W_ES2 ZN_ES2 F_CI2
451      0      0      0      0      0     0     0      0     0
    CL_CI2 NO2_CI2 BR_CI2 NO3_CI2 PO4_CI2 SO4_CI2 CONDU_4 PH
451      0       0      0       0       0       0       0  0
```

Há no conjunto dados faltantes de forma que o número de dados disponíveis para cada

variável individualmente pode diferir do total de pontos disponíveis (697), conforme a seguir.

AG_ES2	AL_ES2	B_ES2	BA_ES2	CA_ES2	CD_ES2	CO_ES2	CR_ES2
697	697	697	697	697	697	697	697
CU_ES2	FE_ES2	GA_ES2	IN_ES2	K_ES2	LI_ES2	MG_ES2	MN_ES2
697	697	652	614	697	697	697	697
MO_ES2	NA_ES2	NI_ES2	PB_ES2	SR_ES2	TL_ES2	V_ES2	W_ES2
697	697	697	697	697	652	697	697
ZN_ES2	F_CI2	CL_CI2	NO2_CI2	BR_CI2	NO3_CI2	PO4_CI2	SO4_CI2
697	695	697	541	627	694	240	694
CONDU_4	PH						
697	697						

A primeira fase consiste em caracterizar o comportamento dos dados utilizando ferramentas exploratórias de forma a:

- evidenciar/destacar possíveis problemas nos dados,
- orientar decisões de como utilizar cada uma das variáveis (transformações, etc),
- destacar e decidir sobre dados claramente discrepantes.

Na análise inicial consideramos que as variáveis podem ser separadas em três grandes grupos:

1. **Grupo I:** variáveis com clara espacialização.
2. **Grupo II:** variáveis quel talvez possam ser consideradas na análise mas talvez agrupadas (”presença””ausência”, ou mais grupos) pois apresentam grande número de dados.
3. **Grupo III:** variáveis aparentemente sem potencial para serem análisadas por apresentarem quase a totalizade dos valores iguais.

Nas subseções a seguir são apresentadas analíses exploratórias para cada variável, em cada um dos grupos. Para cada variável são mostradas figuras compostas de quatro gráficos, descritas em sentido horário por: (i) um diagrama das localizações dos dados, onde para cada ponto a cor representa o quartil correspondente ao valor medido no ponto, ordenados em: azul, verde, amarelo e vermelho. (ii) dados *versus* coordenada-Y, (ii) dados *versus* coordenada-X, (iv) histograma dos dados sobreposto pela densidade empírica estimada e marcas indicando posição dos pontos. Para cada variável levantam-se tópicos para discussão e/ou esclarecimento junto à equipe d o projeto.

1.1 Grupo I

As variáveis deste grupo possuem claramente padrão espacial bem como distribuição de valores que permitem adoção de modelos geoestatísticos simples, podendo ser usados diretamente os valores medidos, tipicamenta após alguma transformação e/ou eventual remoção de valores atípicos. De forma geral a transformação logarítmica (neperiano) é indicada.

Fazem parte descr grupo de variáveis os elementos: Calcio (Ca), potássio (K), Magnésio (Mg), Manganês (Mn), Estrôncio (Sr), Cloro (Cl), Bromo (Br), e os compostos: nitrato (NO_3), fosfato (PO_4), sulfato (SO_4) além das variáveis de pH e *condu*.

Cálcio (Ca)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   2.300   3.810   6.006   7.140  48.610
```

Os dados de cálcio mostram uma distribuição claramente assimétrica e a família de transformações Box-Cox sugere a *transformação logarítmica para estes dados*. A Figura fig:ca-ori mostra os dados originais e a Figura fig:ca0 mostra os dados transformados, onde nota-se claramente um dado discrepante. Neste caso o dado original corresponde a 0.1 enquanto que excluído este o mínimo valor registrado passa a ser 0.41.

Potássio (K)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.050   0.600   0.890   1.331   1.320 102.600
```

Inicialmente notas-se no dados originais de potássio (Figura 4) alguns valores extremamente atípicos, 102.6 e 43.6. A Figura 5 mostra os dados na escala logarítmica o que também evidencia um valor discrepante muito pequeno.

As Figuras 6 e 7 mostram dos dados originais e transformados após a remoção destes três valores discrepantes. Os três dados e suas posições são listadas a seguir. *PJ: precisa-se discutir se estes valores são plausíveis, erro de digitação, etc*

```
    BACIAS_ID LABEL_UTME LABEL_UTMN LONGITUDE  LATITUDE  SIGLA
509       536   746823.0    7195483 -48.54778 -25.33684 RL-015
511       540   738946.5    7193118 -48.62556 -25.35946 RL-014
593       629   669059.9    7151187 -49.31446 -25.74751 IG-008
    DATA_ANALI  K_ES2
509 1996-03-08 102.60
511 1996-03-08  43.60
593 1996-03-19   0.05
```

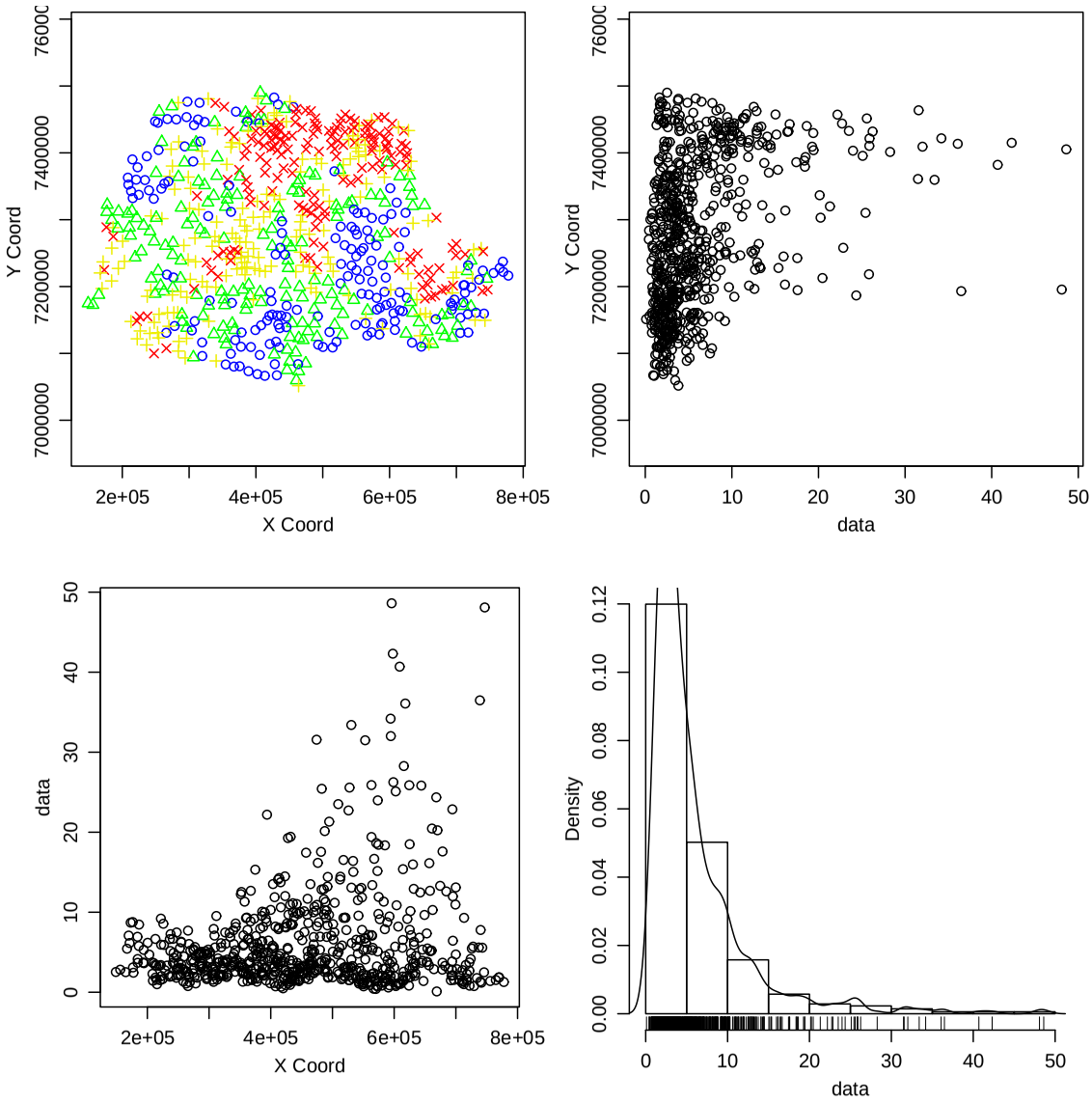

Figure 1: Cálcio (Ca), dados originais

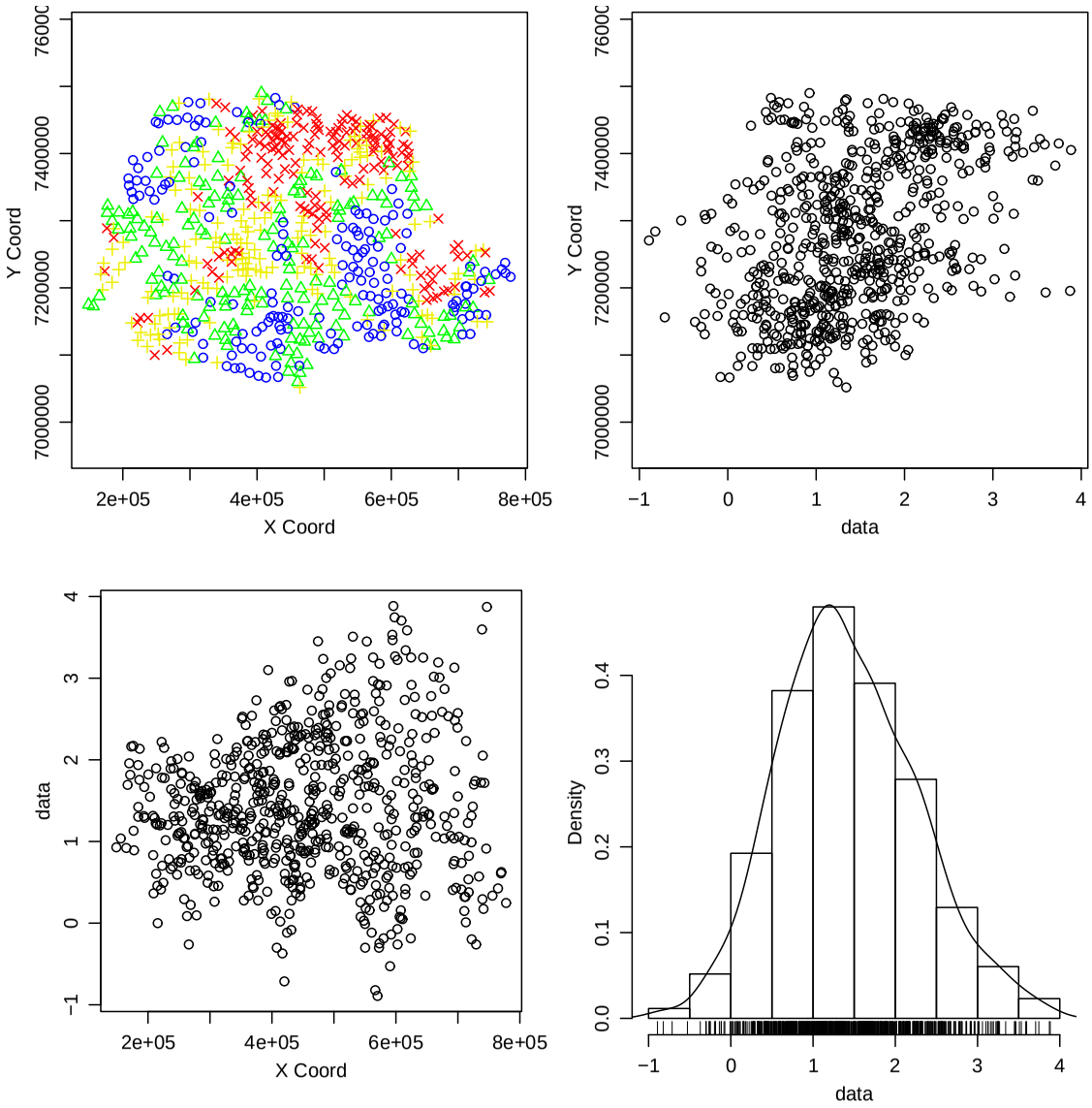

Figure 2: Cálcio (Ca), dados transformados (logarítmo)

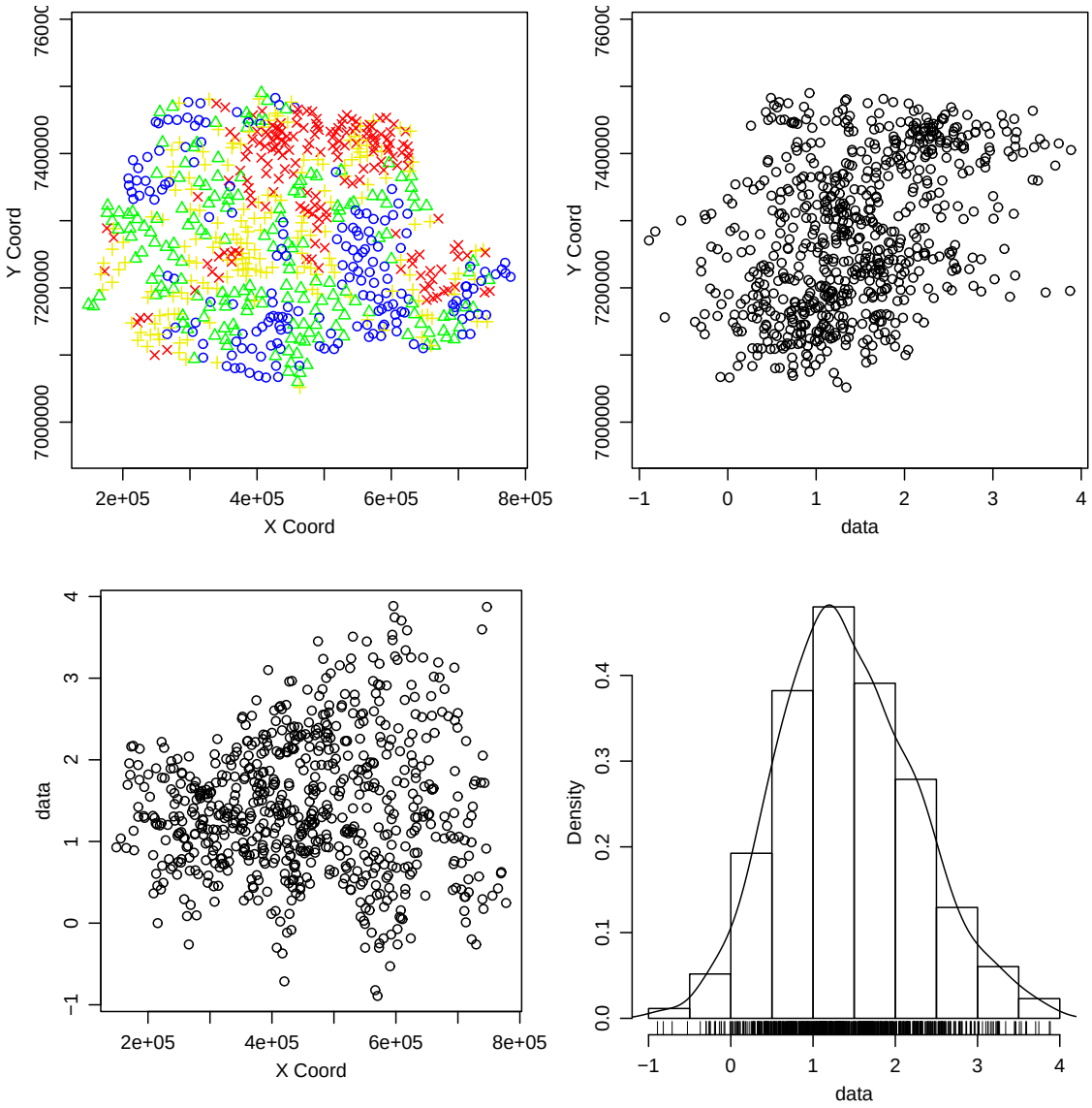

Figure 3: Cálcio (Ca), dados transformados (logarítmo), excluindo o dado discrepante.

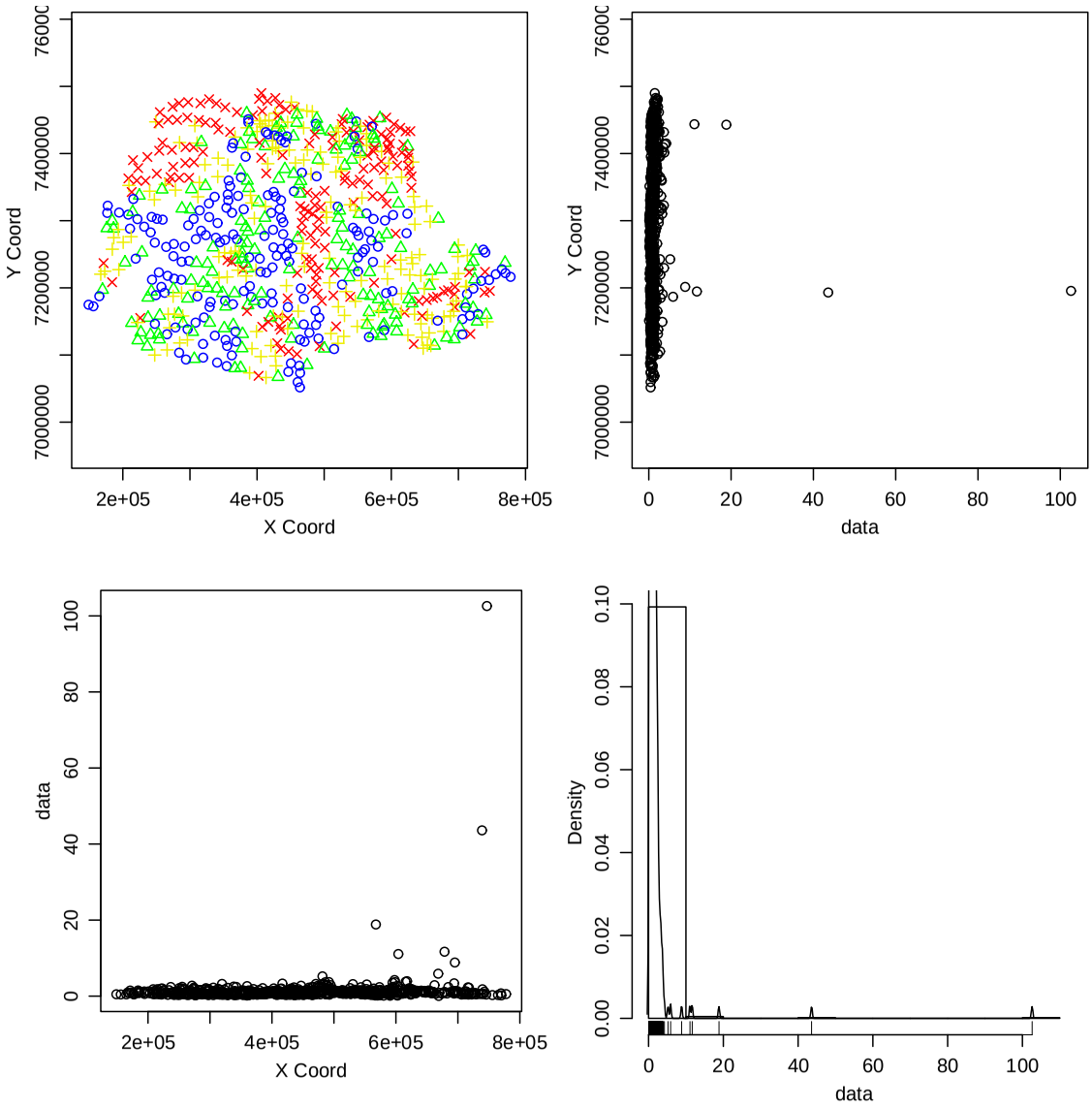

Figure 4: Potássio (K), dados originais.

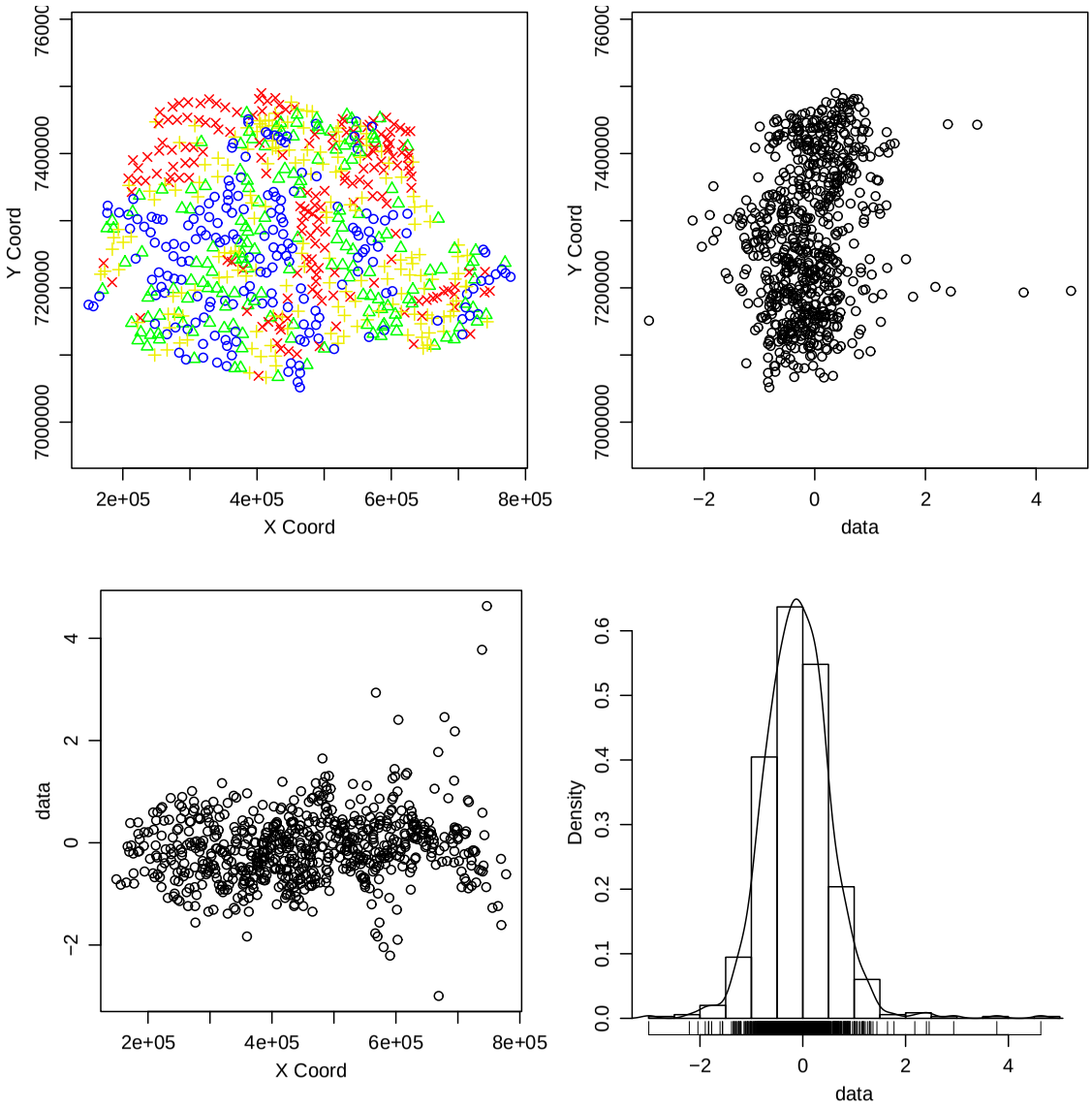

Figure 5: Potássio (K), dados transformados (log).

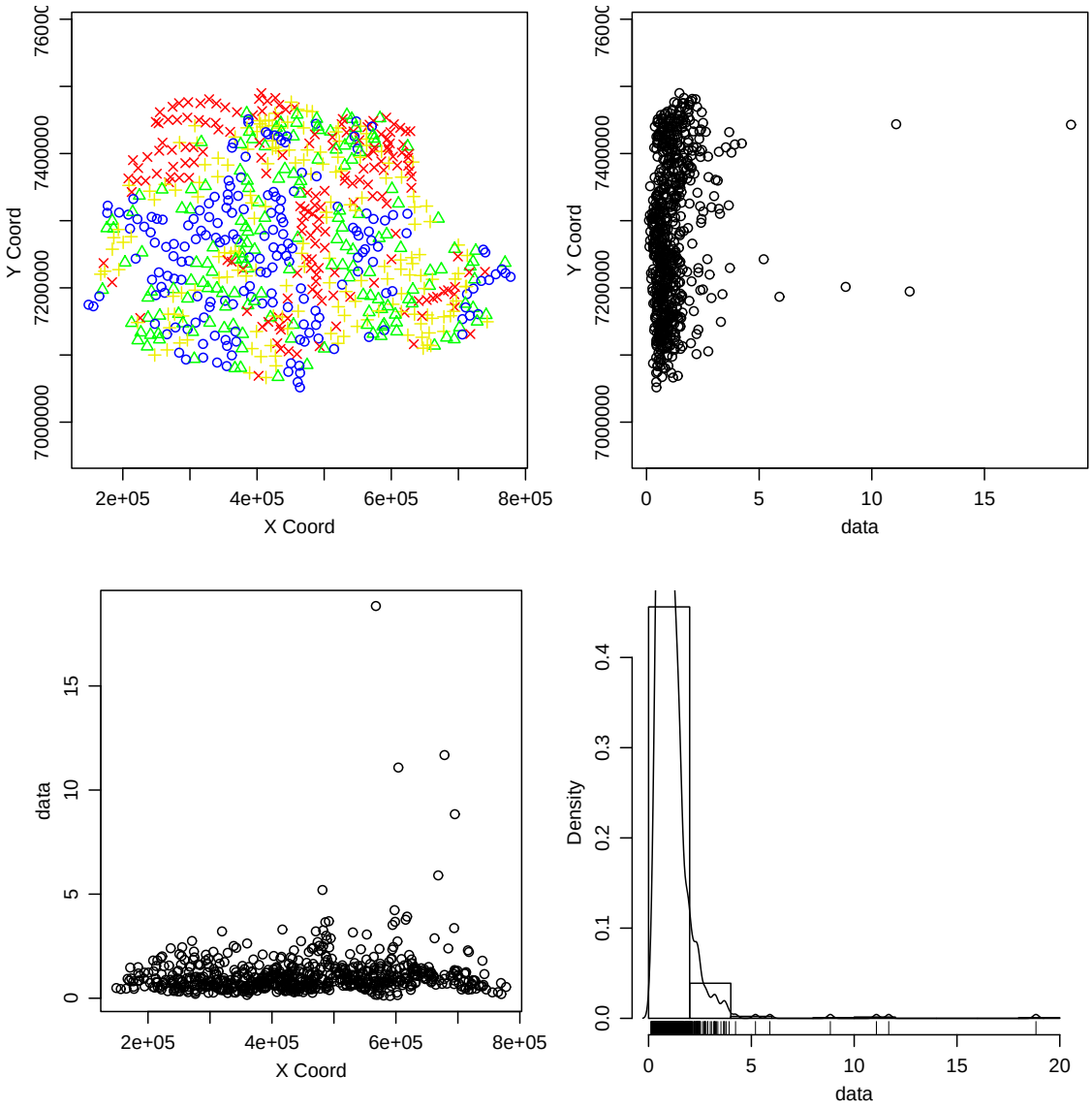

Figure 6: Potássio (K), dados originais, excluindo três valores discrepantes.

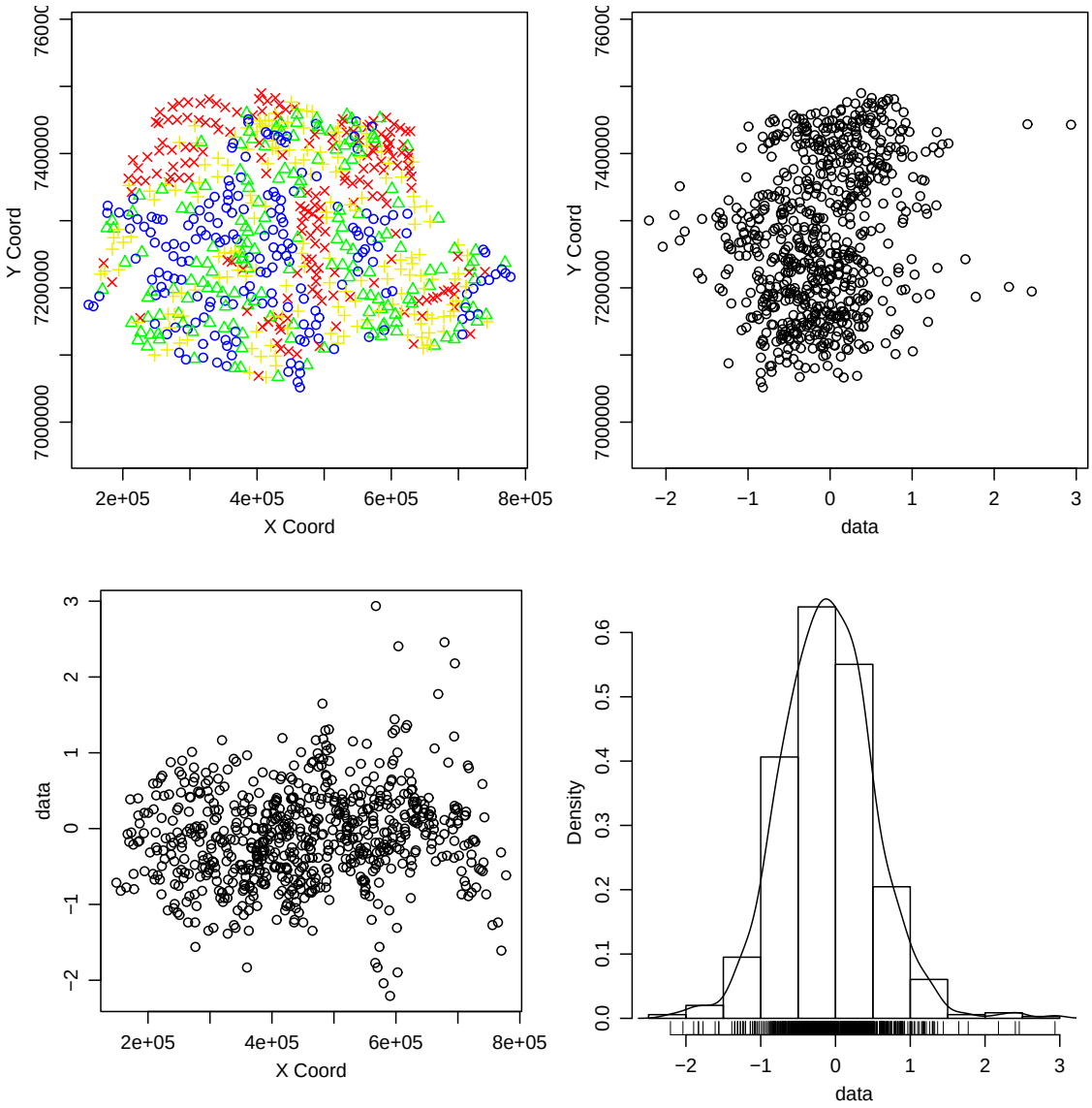

Figure 7: Potássio (K), dados transformados (log), excluindo três valores discrepantes.

Magnésio (Mg)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.025   1.210   1.810   3.020   3.280 159.300
```

Sódio (Na)

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.125    1.010    1.680    5.672    2.830 1244.000
```

Estrôncio (Sr)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00600 0.02000 0.03000 0.04275 0.05000 0.70000
```

Cloro (Cl)

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.008    0.430    0.800    7.721    1.500 2530.000
```

Nitrato (N03)

```
as.geodata: 3 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.010   0.580   1.300   2.032   2.500  32.400
```

Fosfato (PO4)

```
as.geodata: 457 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01000 0.01000 0.02000 0.04667 0.04000 0.83000
```

Permanece assimétrico no logarítimo. Ver limites nas medidas

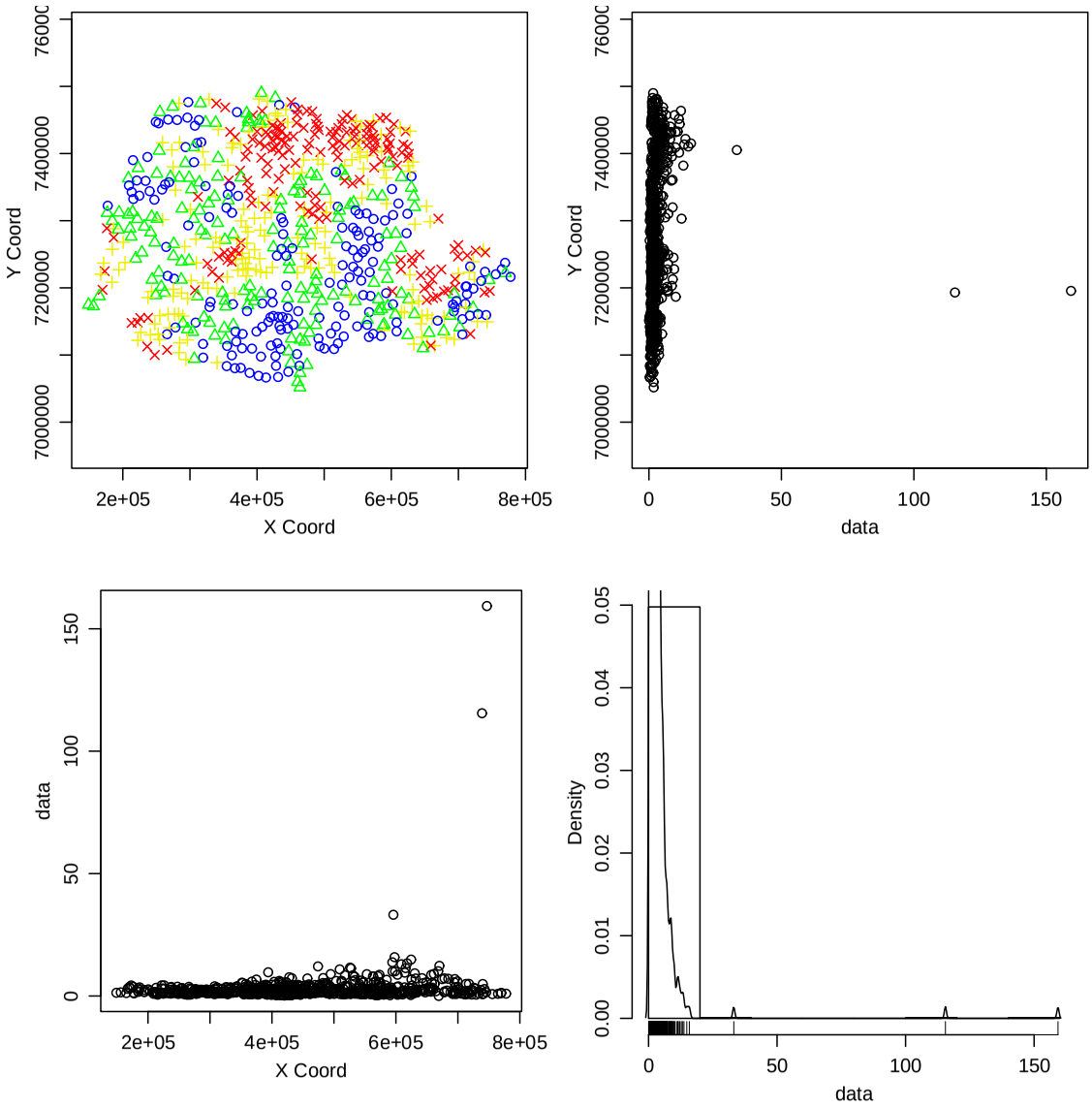

Figure 8: Magnésio (Mg), dados originais.

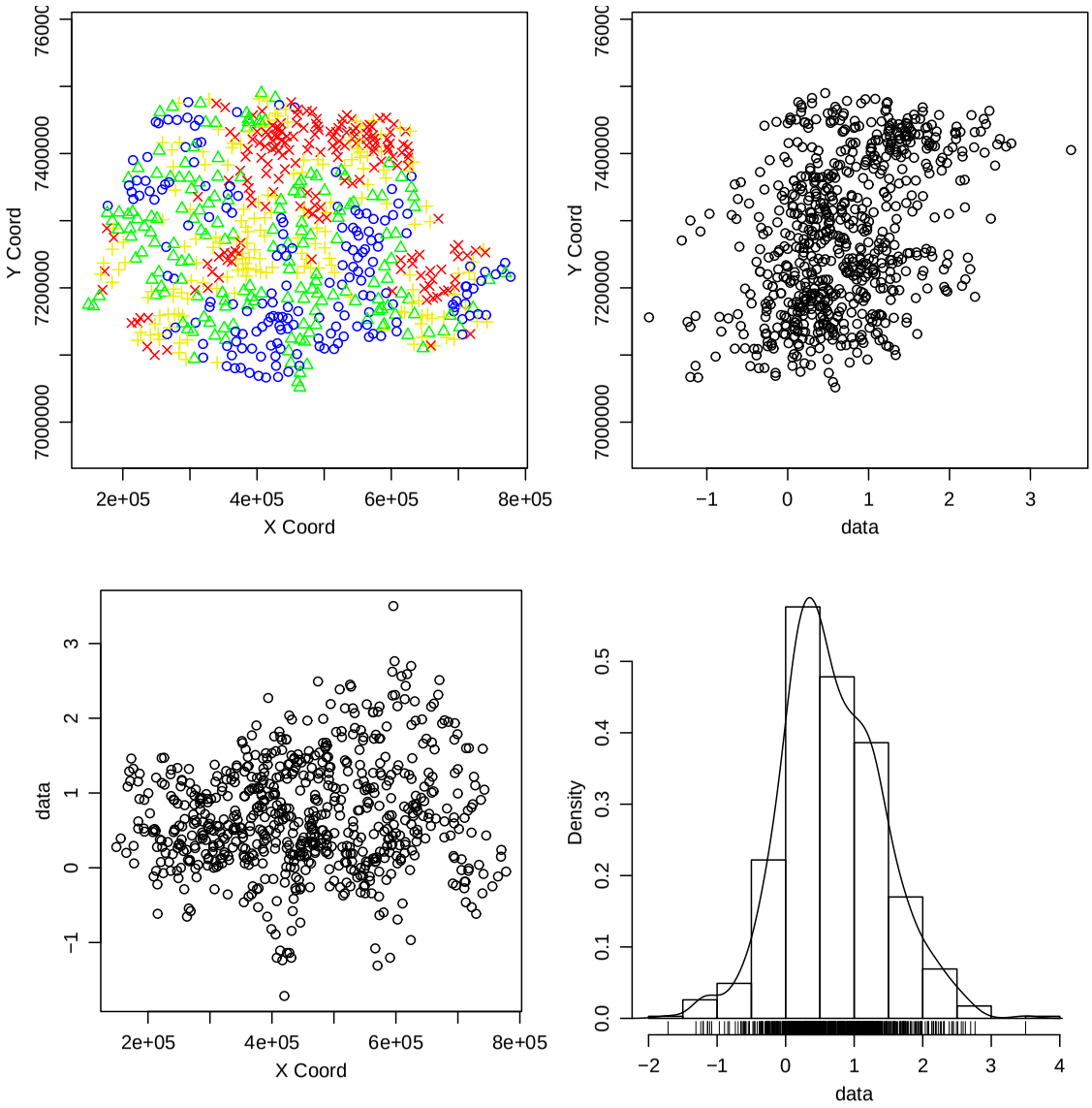

Figure 9: Magnésio (Mg), dados transformados (log).

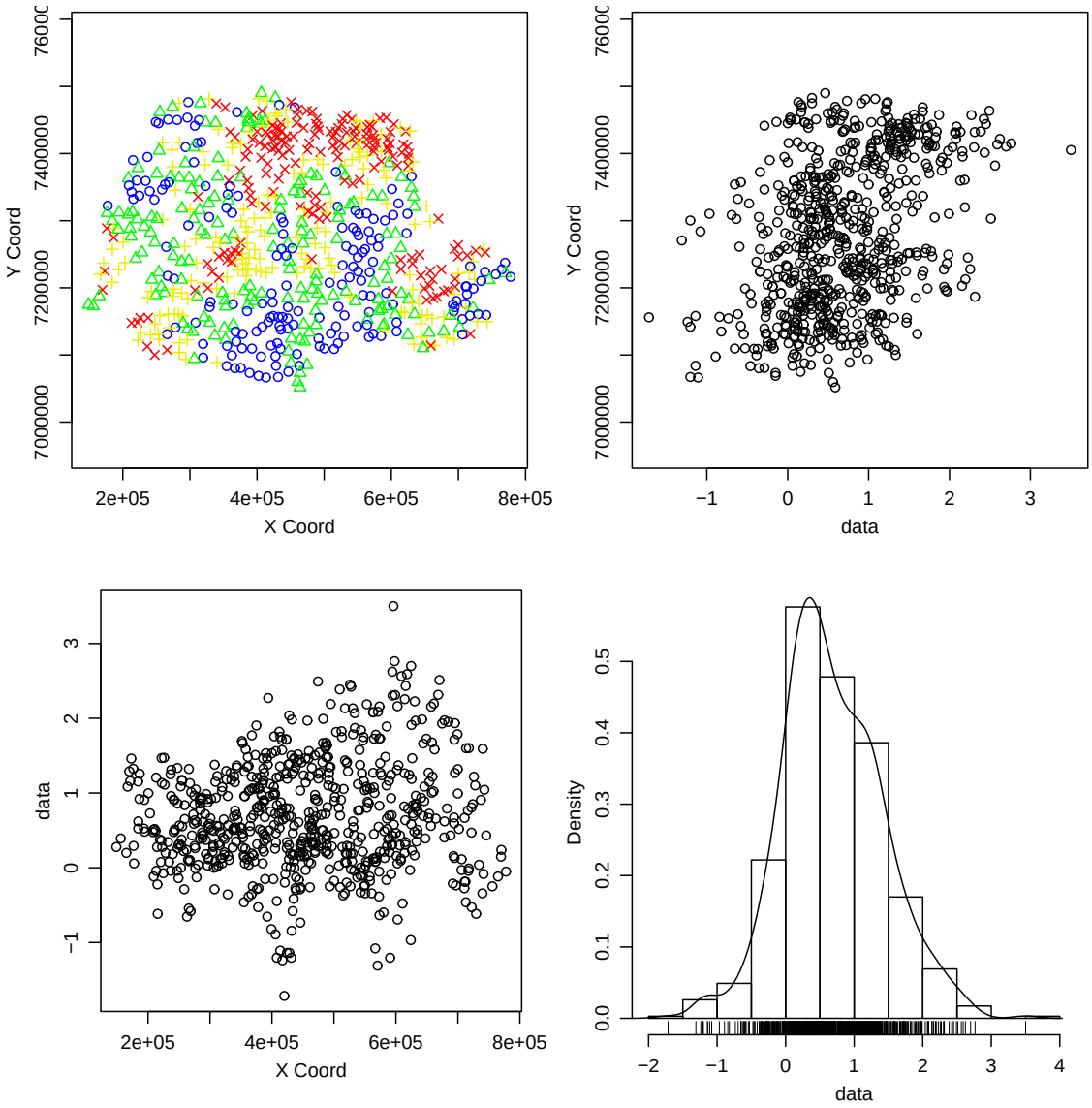

Figure 10: Magnésio (Mg), retirados três dados discrepantes e transformados (log).

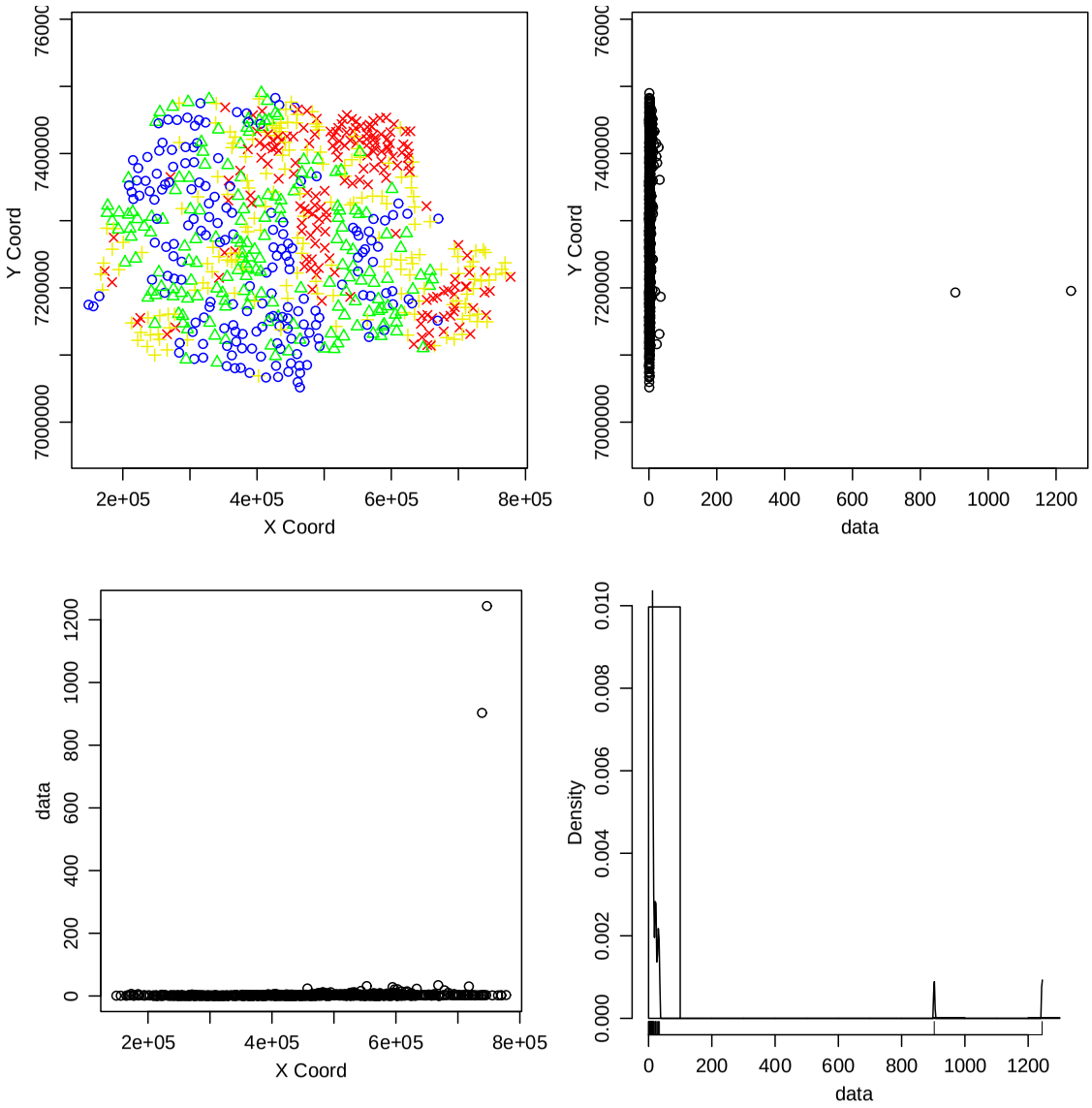

Figure 11: Sódio (Na), dados originais.

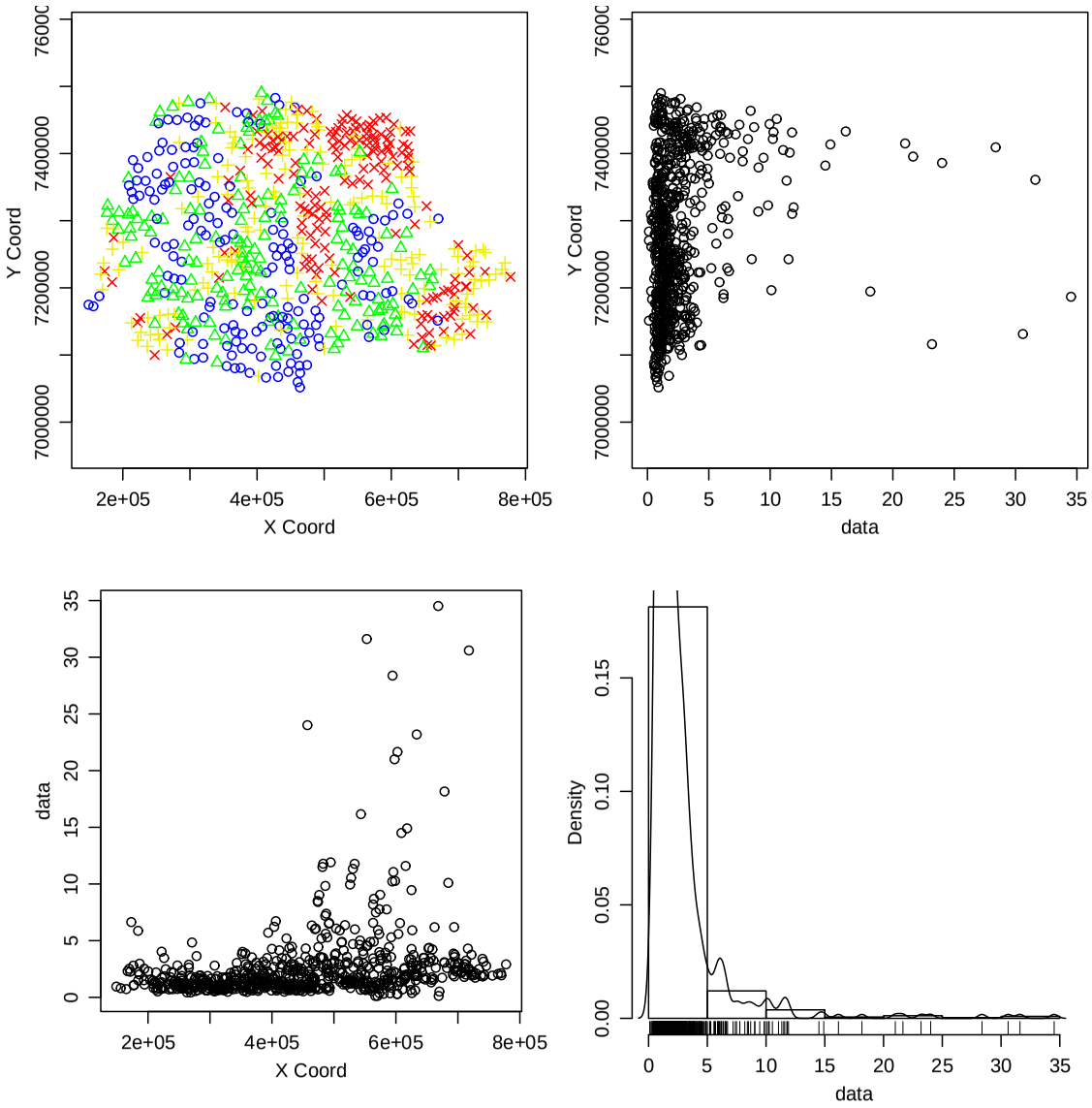

Figure 12: Sódio (Na), retirados dados > 200.

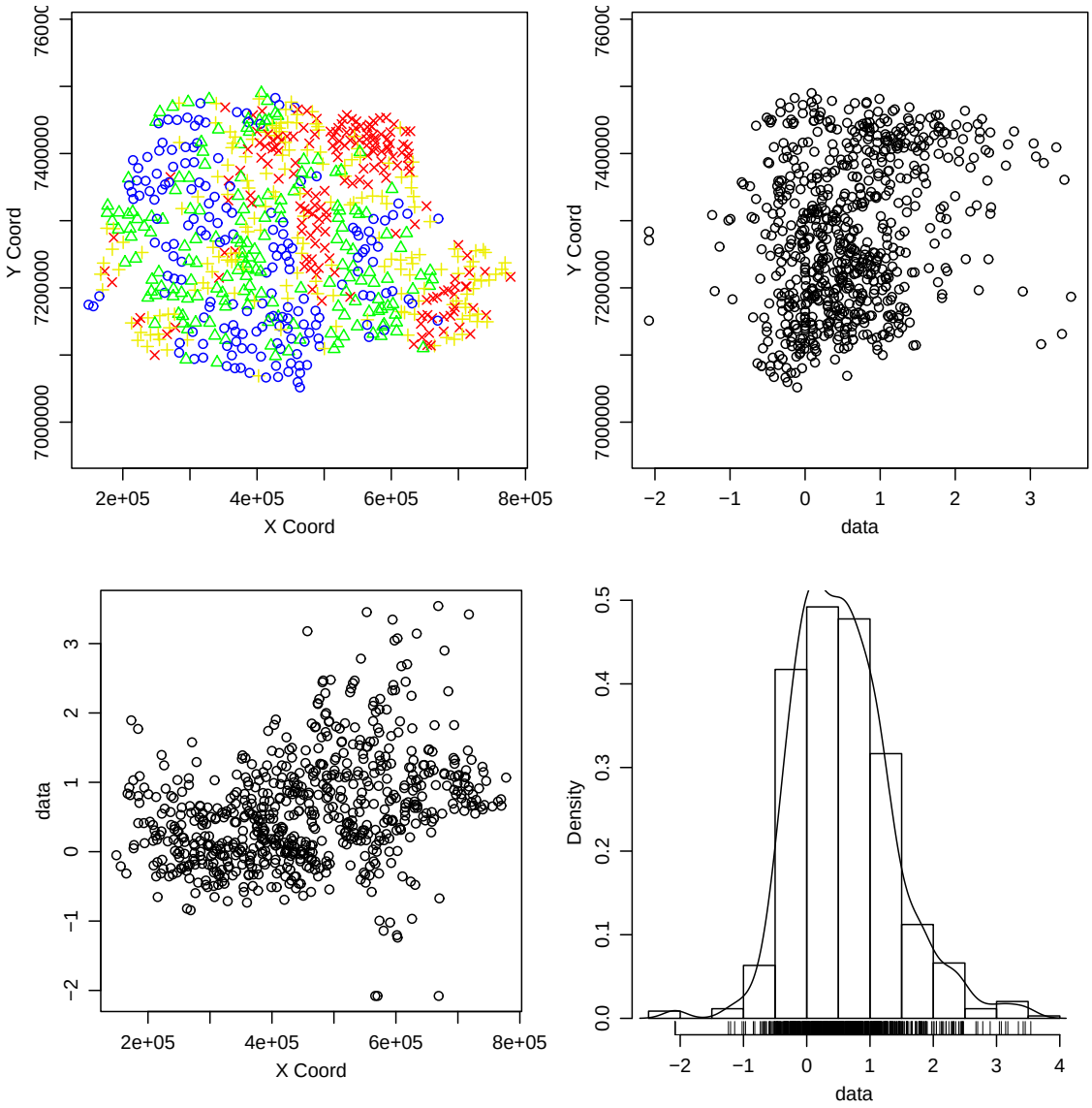

Figure 13: Sódio (Na), retirados dados > 200, transformados (log).

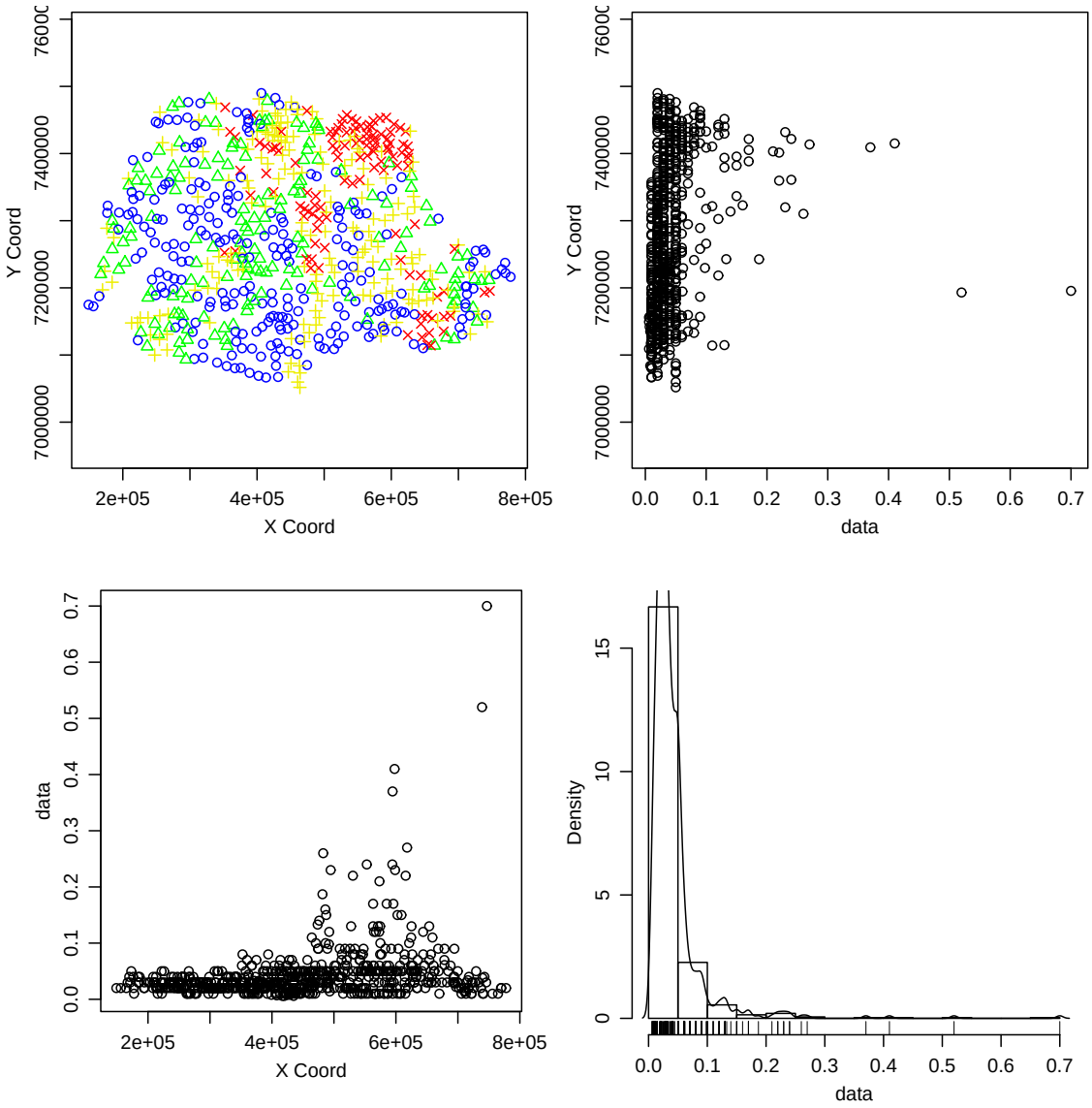

Figure 14: Estrôncio (Sr), dados originais.

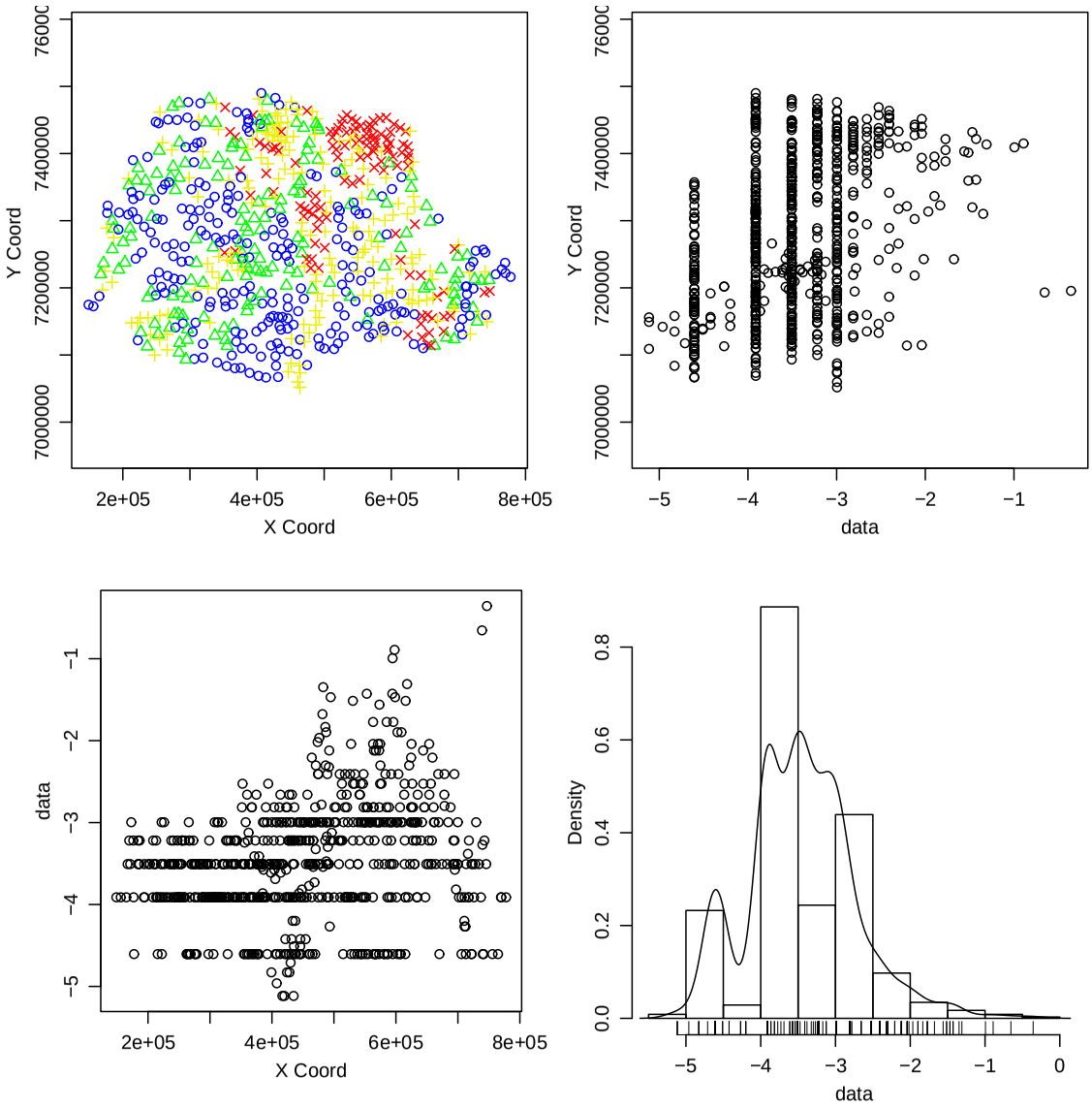

Figure 15: Estrôncio (Sr), dados transformados.

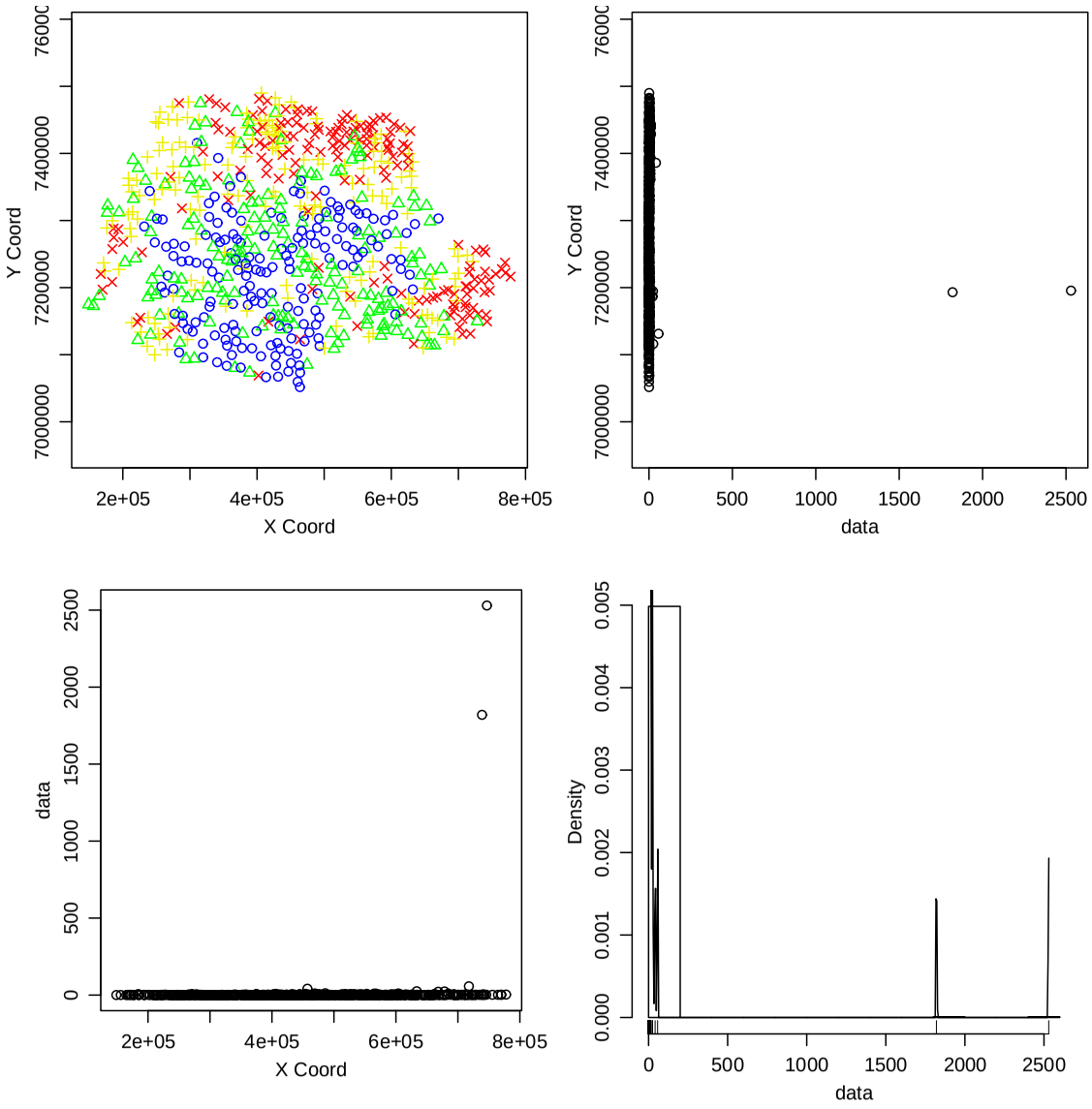

Figure 16: Cloro (Cl), dados originais.

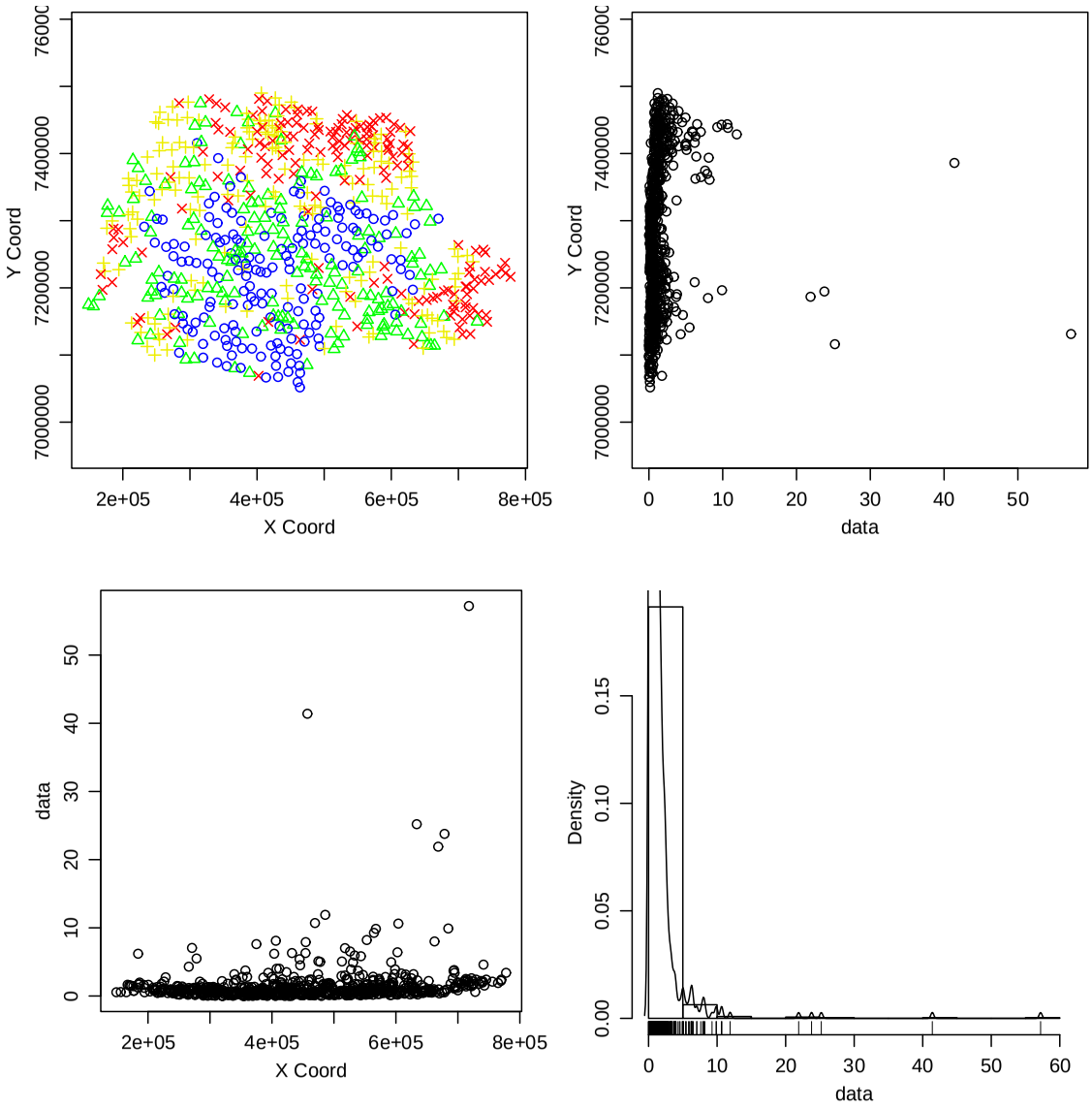

Figure 17: Cloro (Cl), retirados dados > 200.

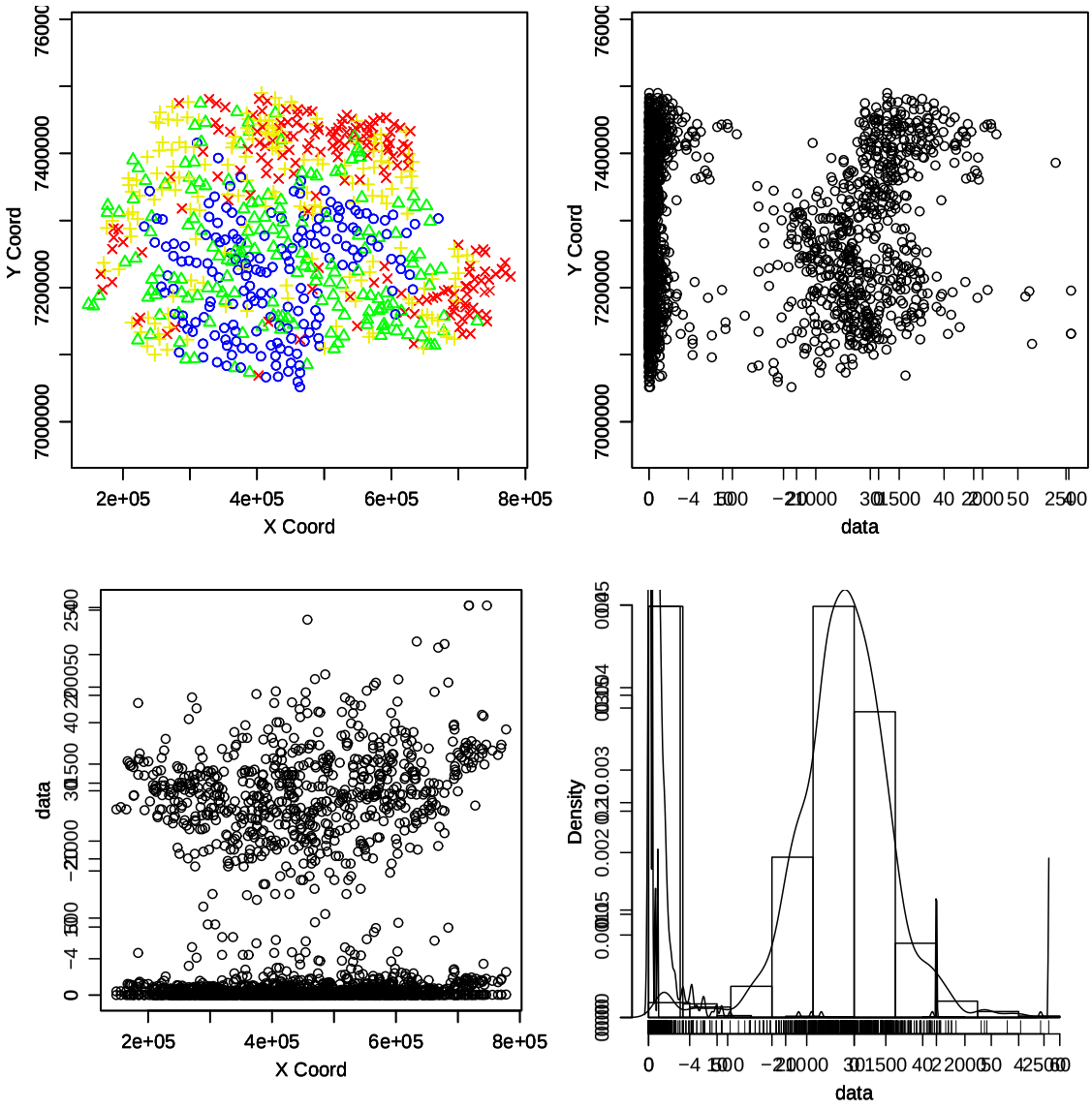

Figure 18: Cloro (Cl), retirados dados > 200 dados transformados (log).

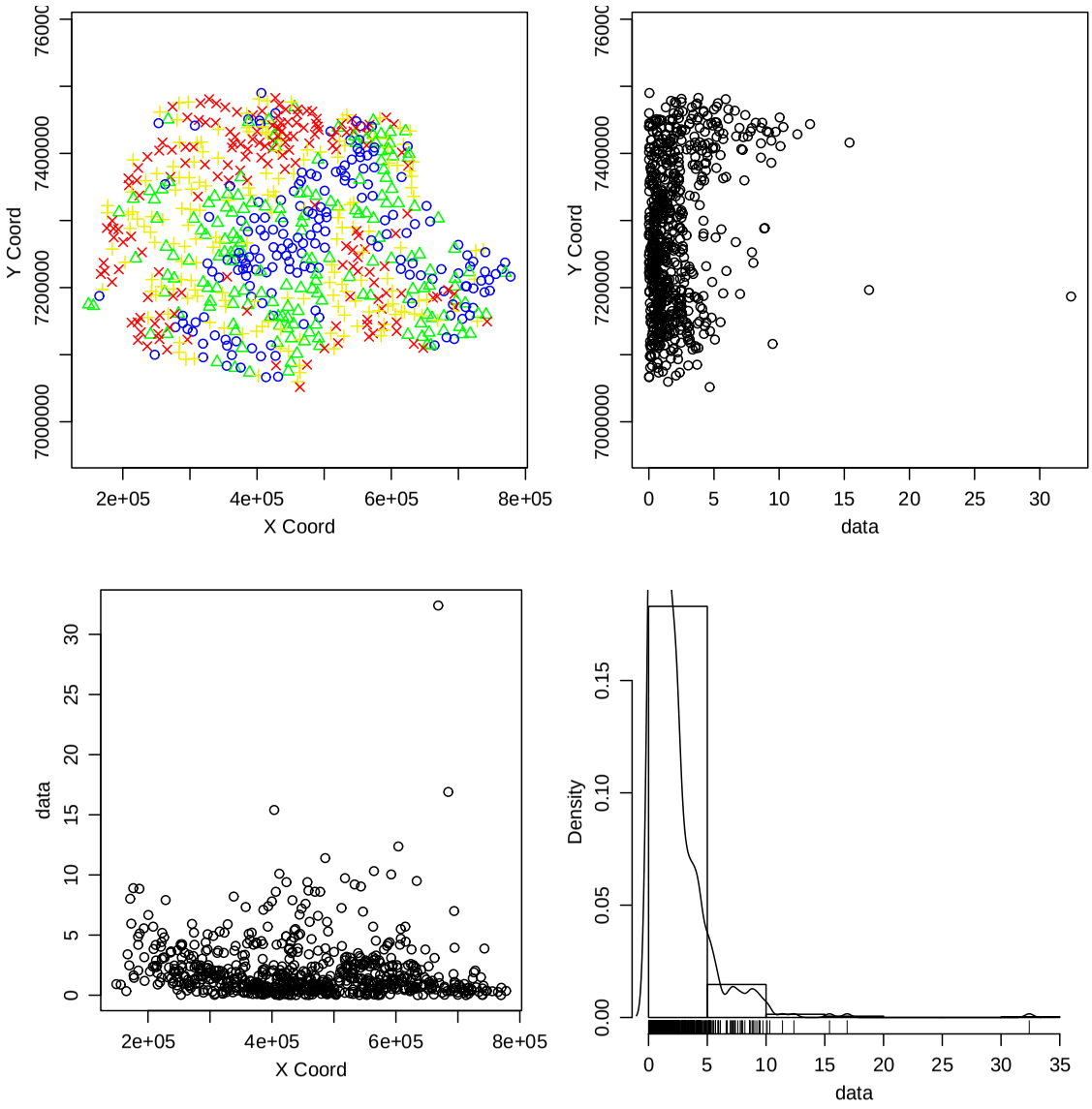

Figure 19: Nitrato (NO_3), dados originais.

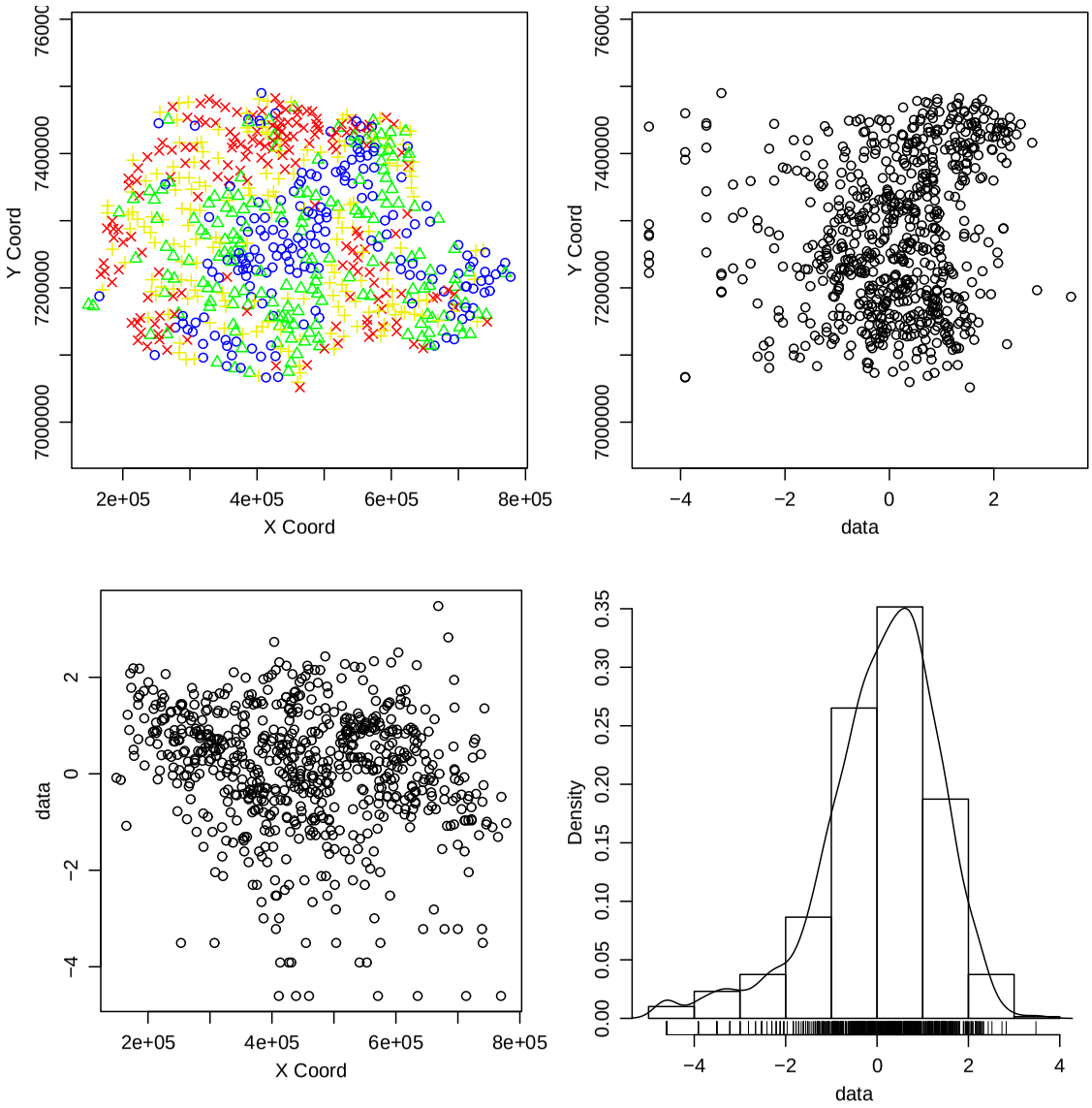

Figure 20: Nitrato (NO_3), transformados (log).

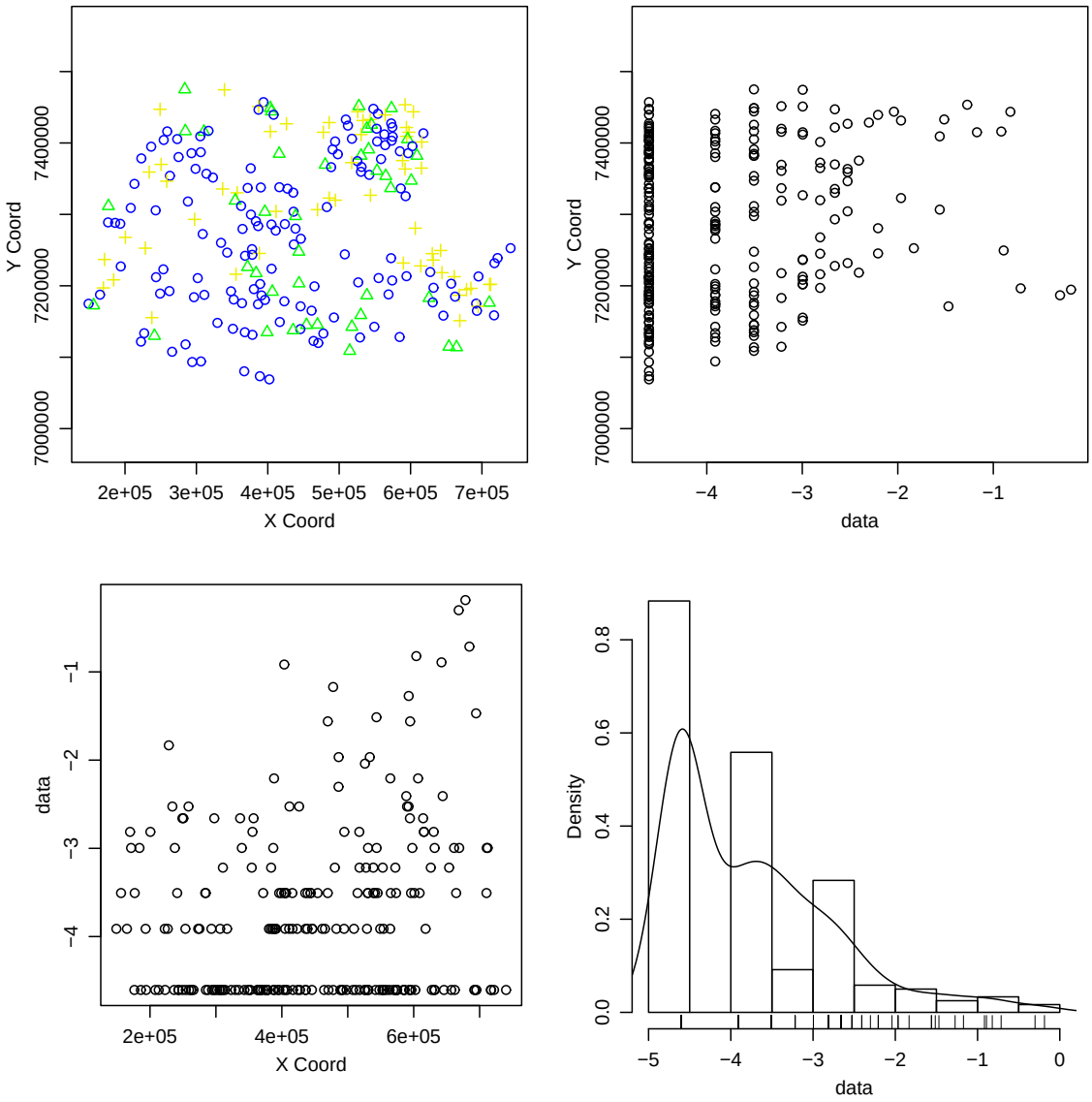

Figure 21: Fosfato (PO_4), dados originais.

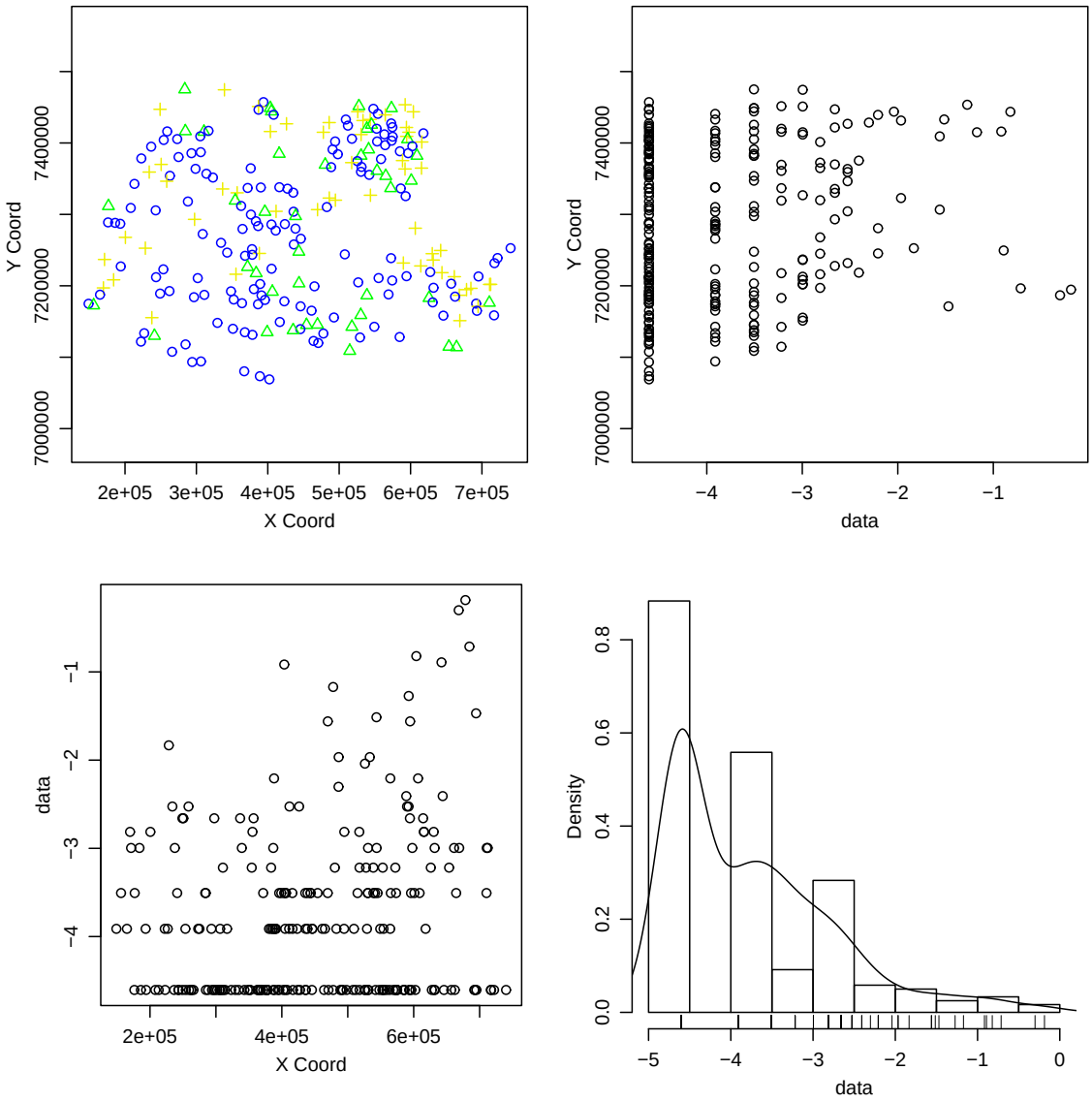

Figure 22: Fosfato (PO_4), dados transformados (log).

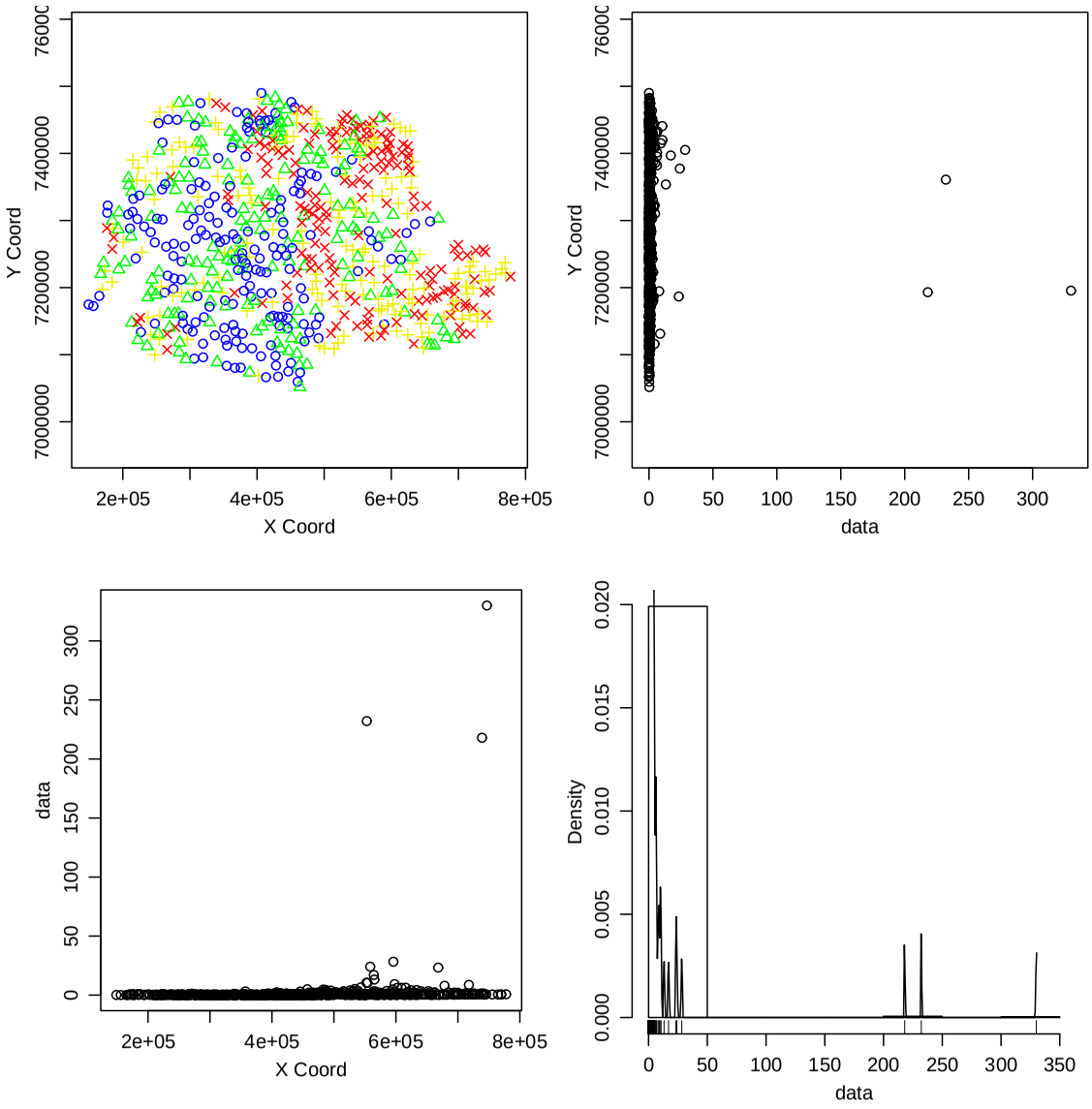

Figure 23: Sulfato (SO_4), dados originais.

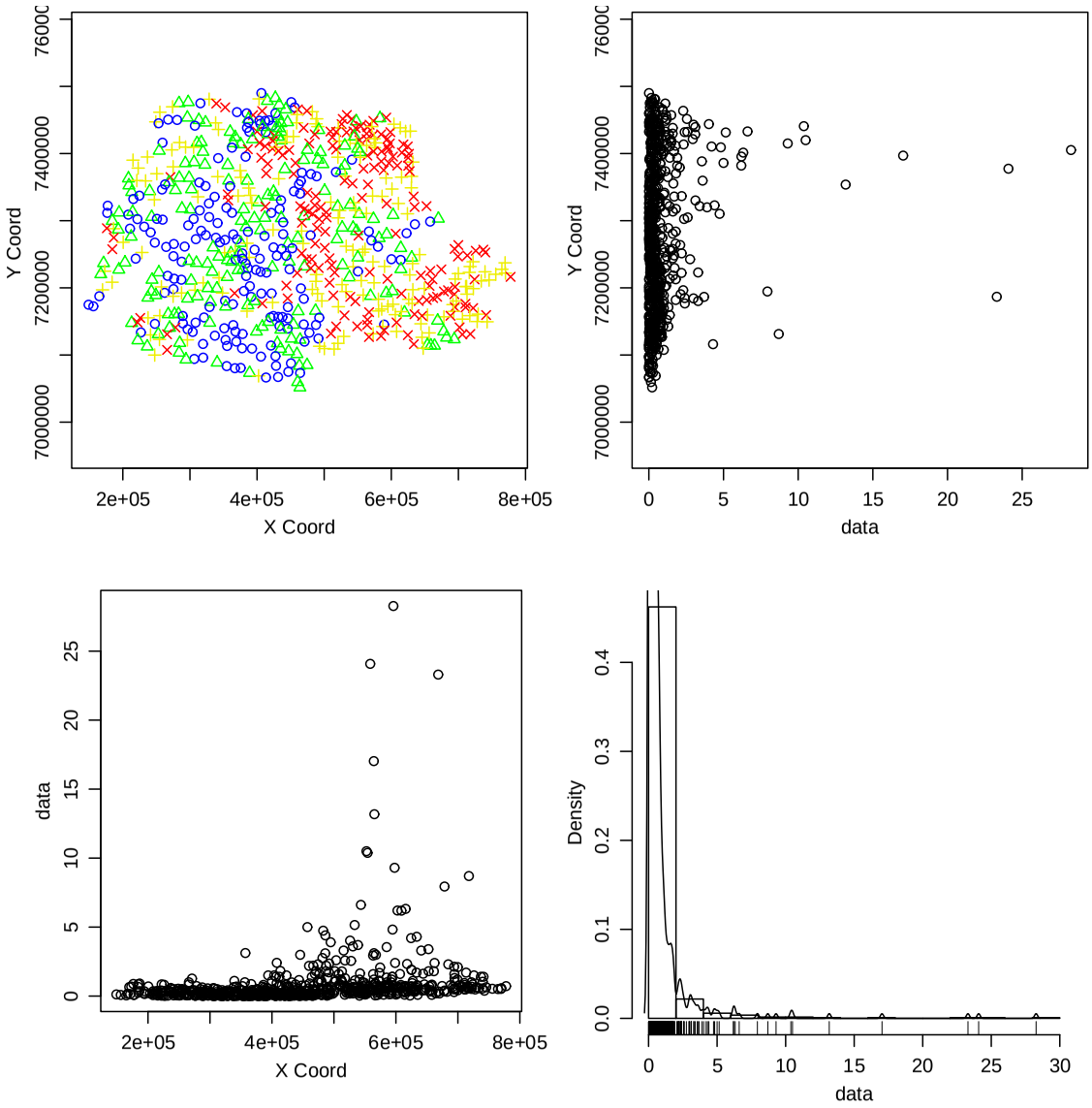

Figure 24: Sulfato (SO_4), retirados dados > 50.

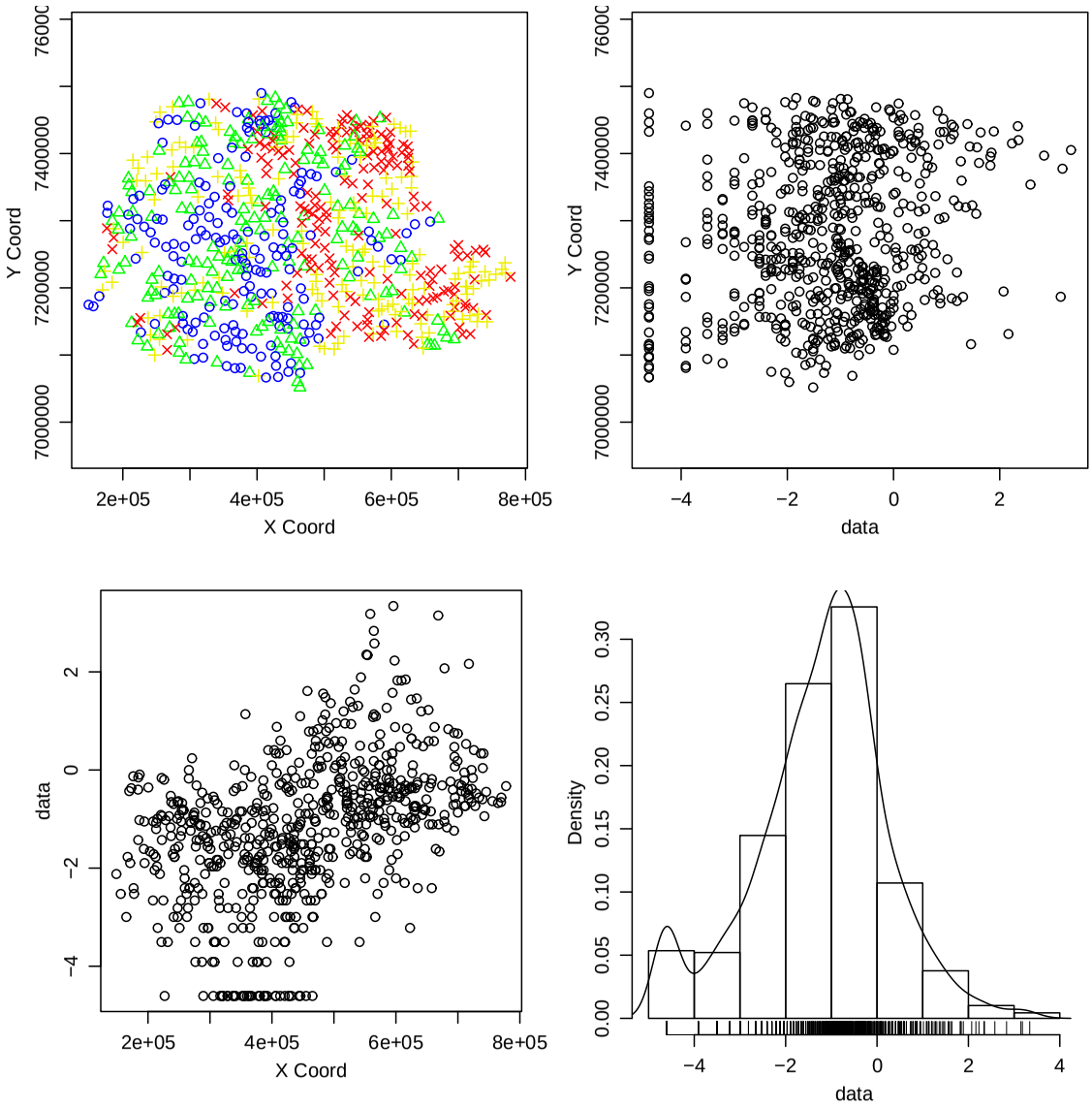

Figure 25: Sulfato (SO_4), retirados dados > 50, transformados (log) .

Sulfato (SO4)

```
as.geodata: 3 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.010   0.140   0.350   1.945   0.710 330.000
```

Acidez (pH)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.300   6.300   6.700   6.591   6.900   7.700
```

Condutibilidade (condu)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.40   30.80   45.10   83.27   79.50 7540.00
```

FAZER UMA TABELA MOSTRANDOS POSIÇOES DOS DADOS ATÍPICOS

1.2 Grupo II

Ferro (Fe)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.0100  0.0400  0.1054  0.1300  2.3500
```

Nota-se que para este elemento o valor mínimo é de 0.01 e registrado em 257 pontos Isto sugere que esta valor representa na verdade o limite mínimo de detecção, o que na literatura estatística é tratado pelo termo *censura à esquerda*.

Outro aspecto que se nota no gráfico é que apenas 3 das quatro cores dos quartis aparecem. Isto ocorre porque o número de dados no limite de deteção é muito alto e o valor mínimo é igual ao do primeiro quartil.

Manganês (Mn)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00100 0.01000 0.01000 0.02555 0.01000 1.44000
```

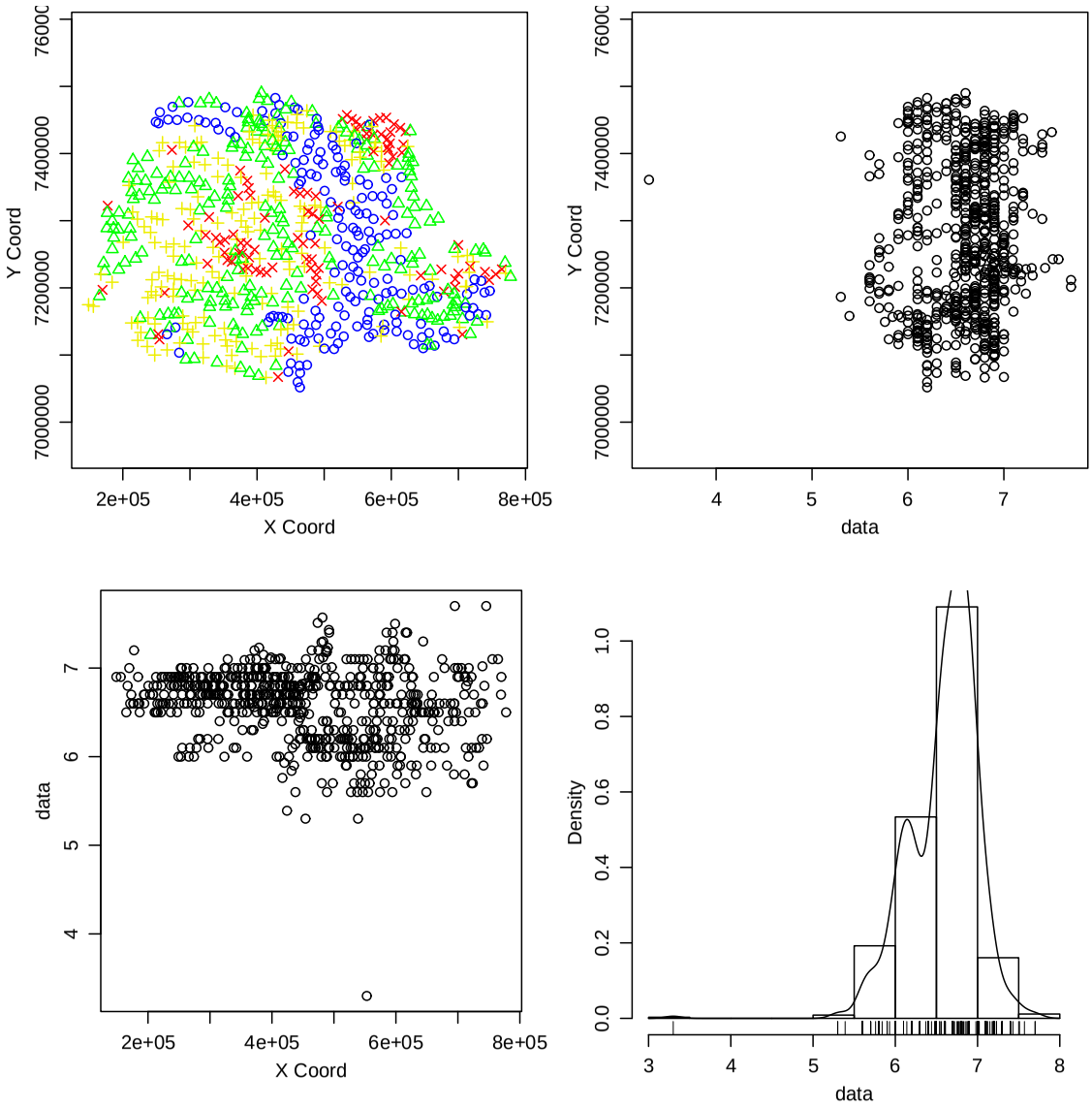

Figure 26: Acidez (pH), dados originais.

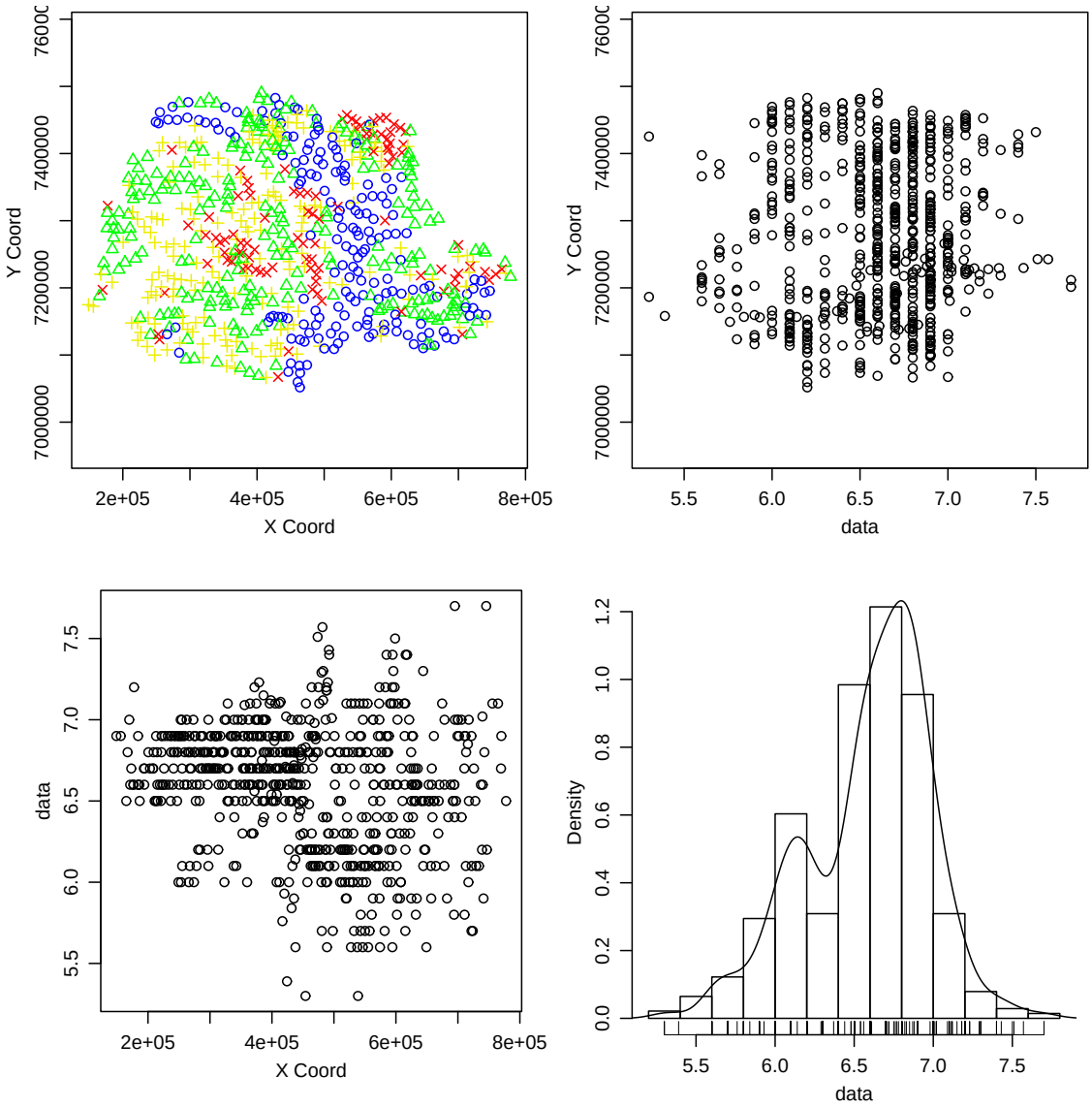

Figure 27: Acidez (pH), dados originais após retirada de valores < 4.

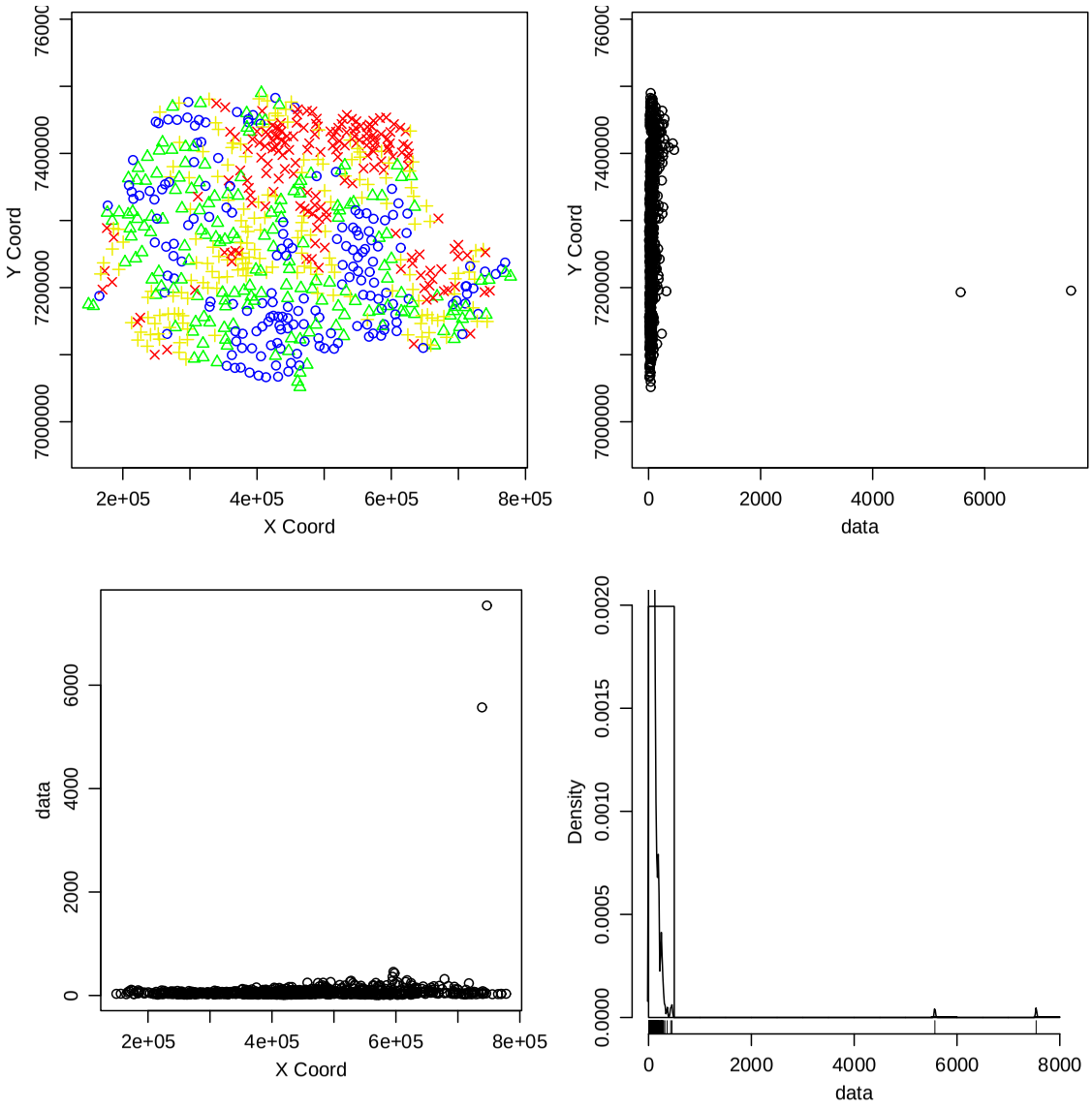

Figure 28: Condutibilidade (Condu), dados originais.

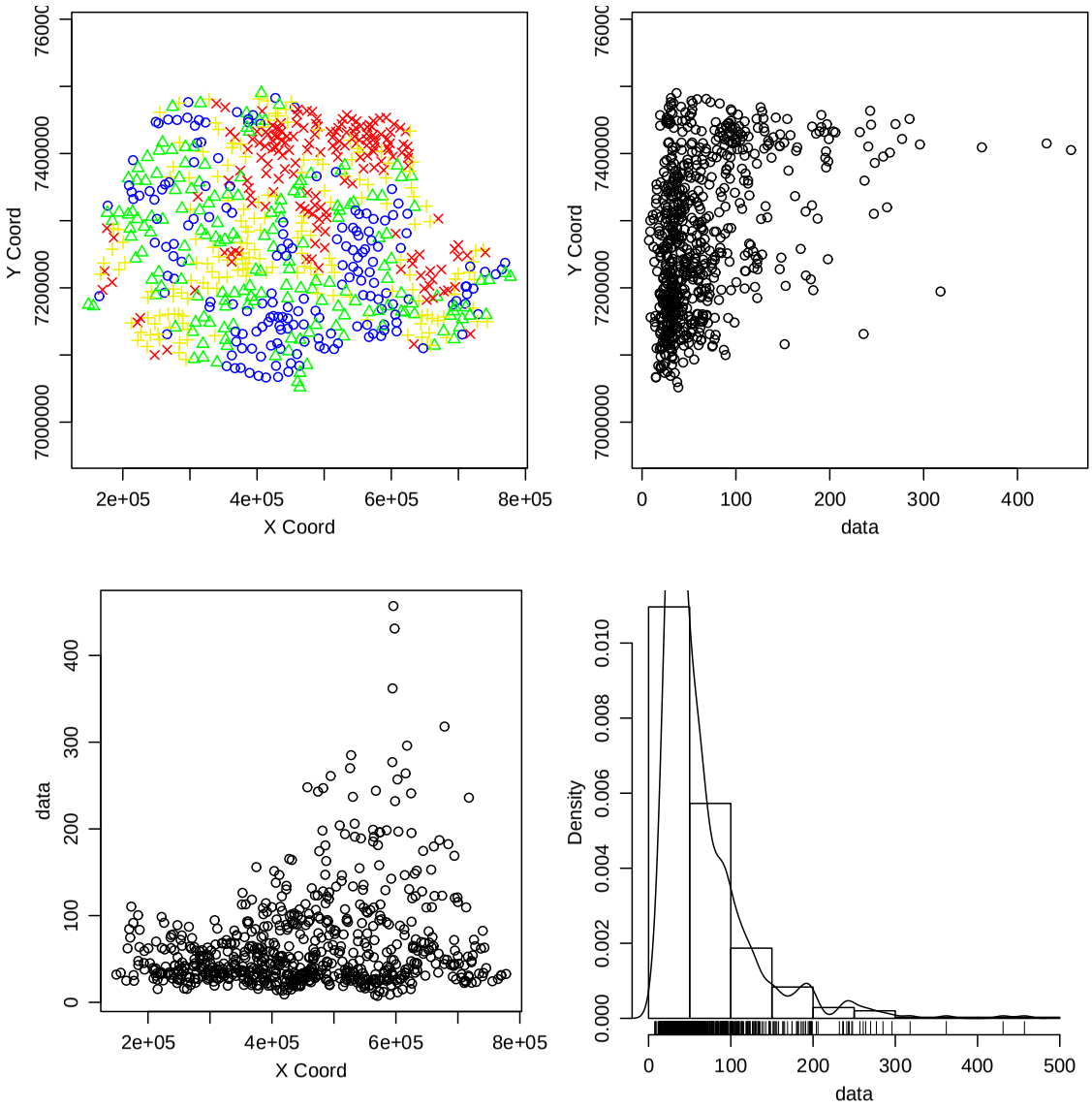

Figure 29: Condutibilidade (condu), retirados dados acima do valor 1000.

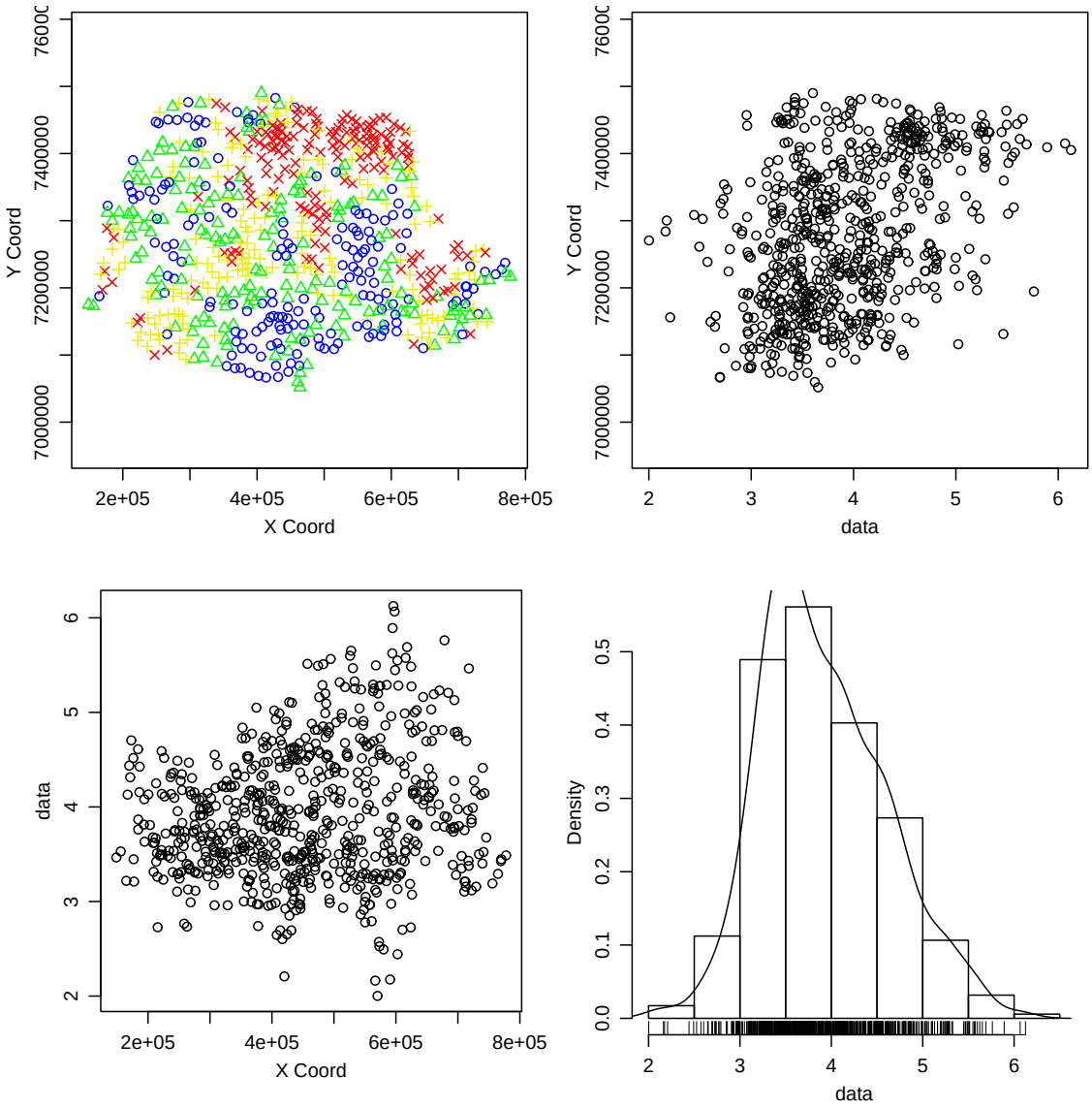

Figure 30: Condutibilidade, retirados dados acima do valor 1000 e transformados (log).

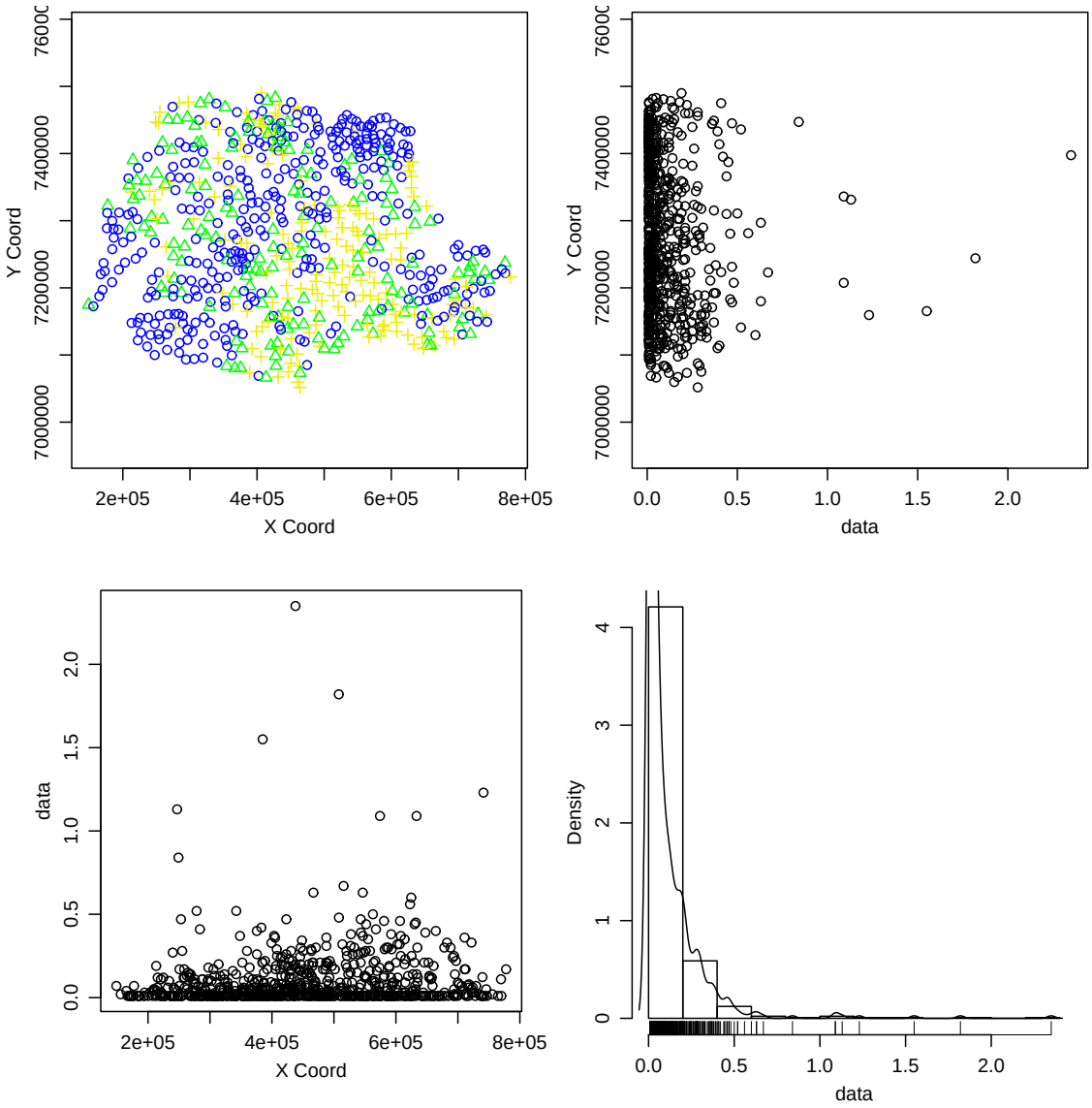

Figure 31: Ferro (Fe), dados originais

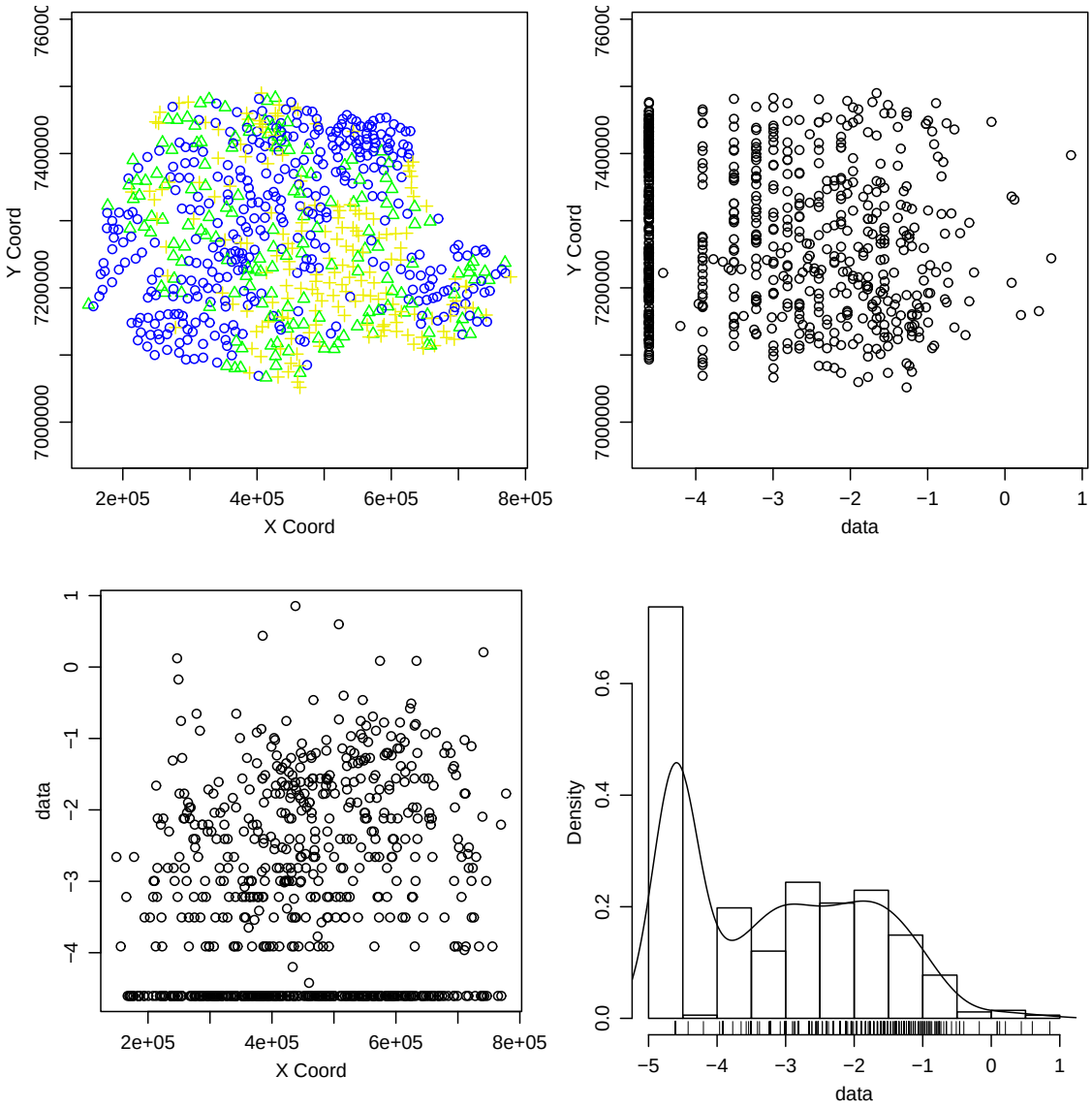

Figure 32: Ferro (Fe), dados transformados (logarítmo).

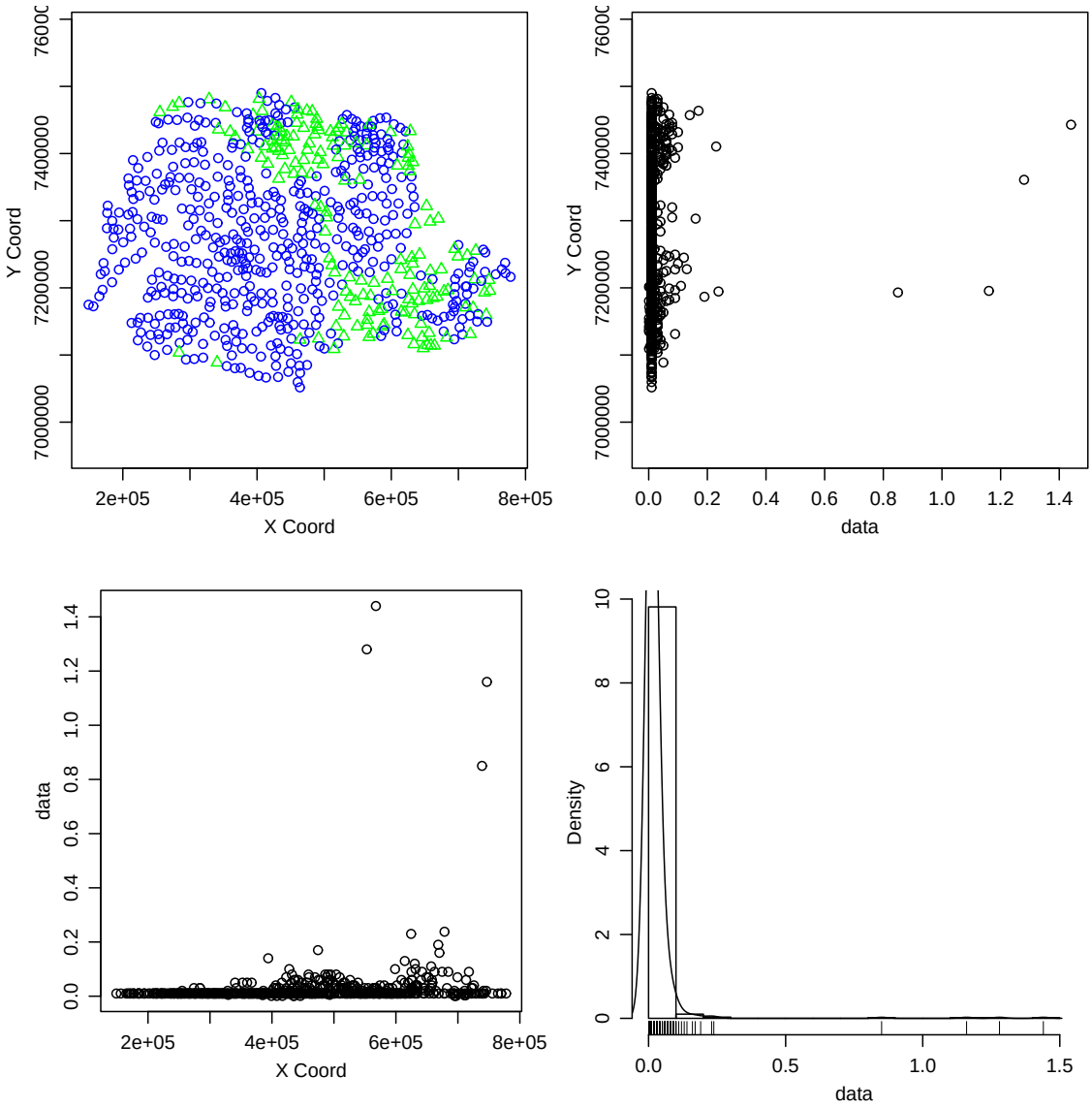

Figure 33: Manganês (Mn), dados originais.

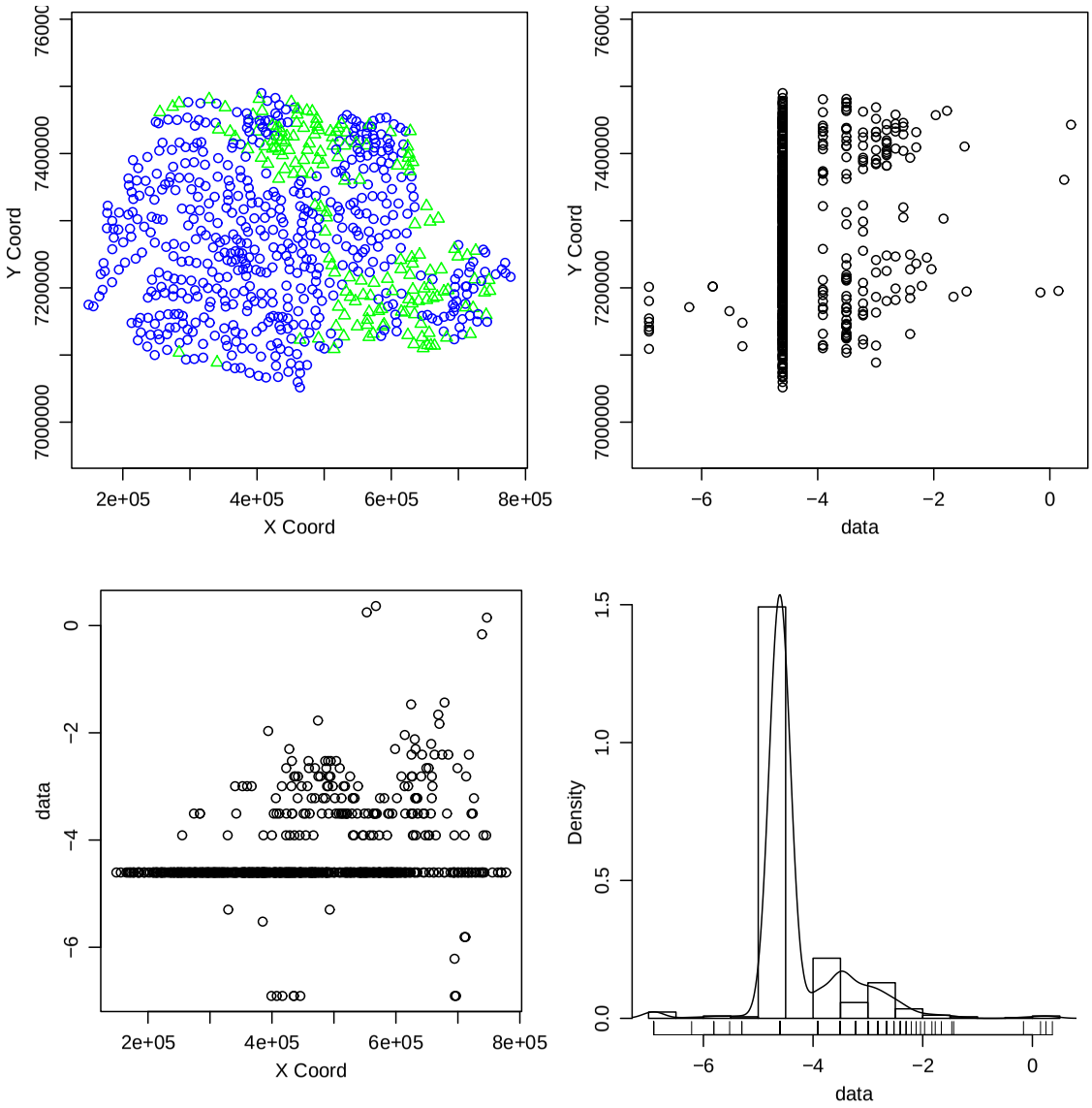

Figure 34: Manganês (Mn), dados originais.

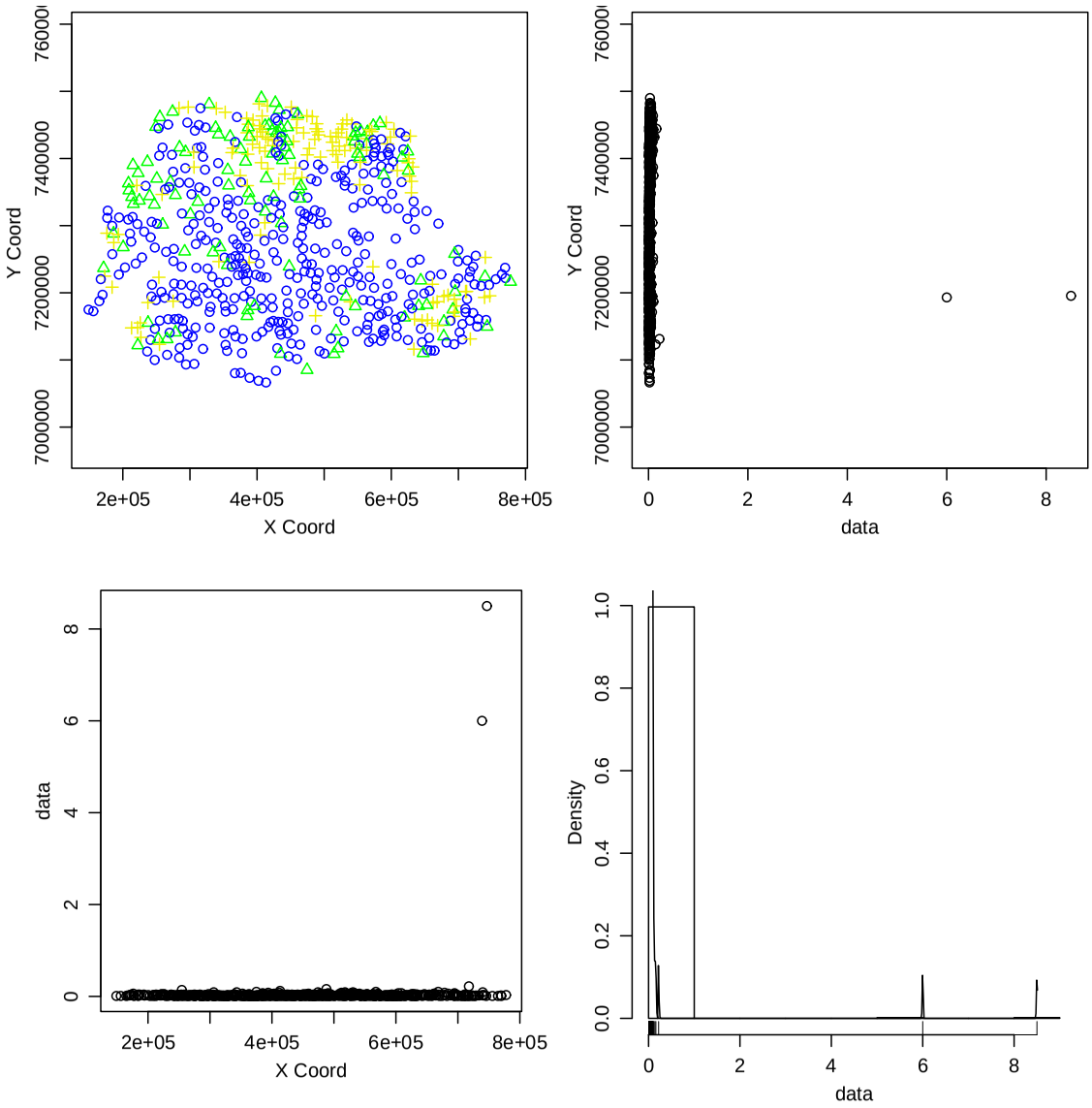

Figure 35: Bromo (Br), dados originais.

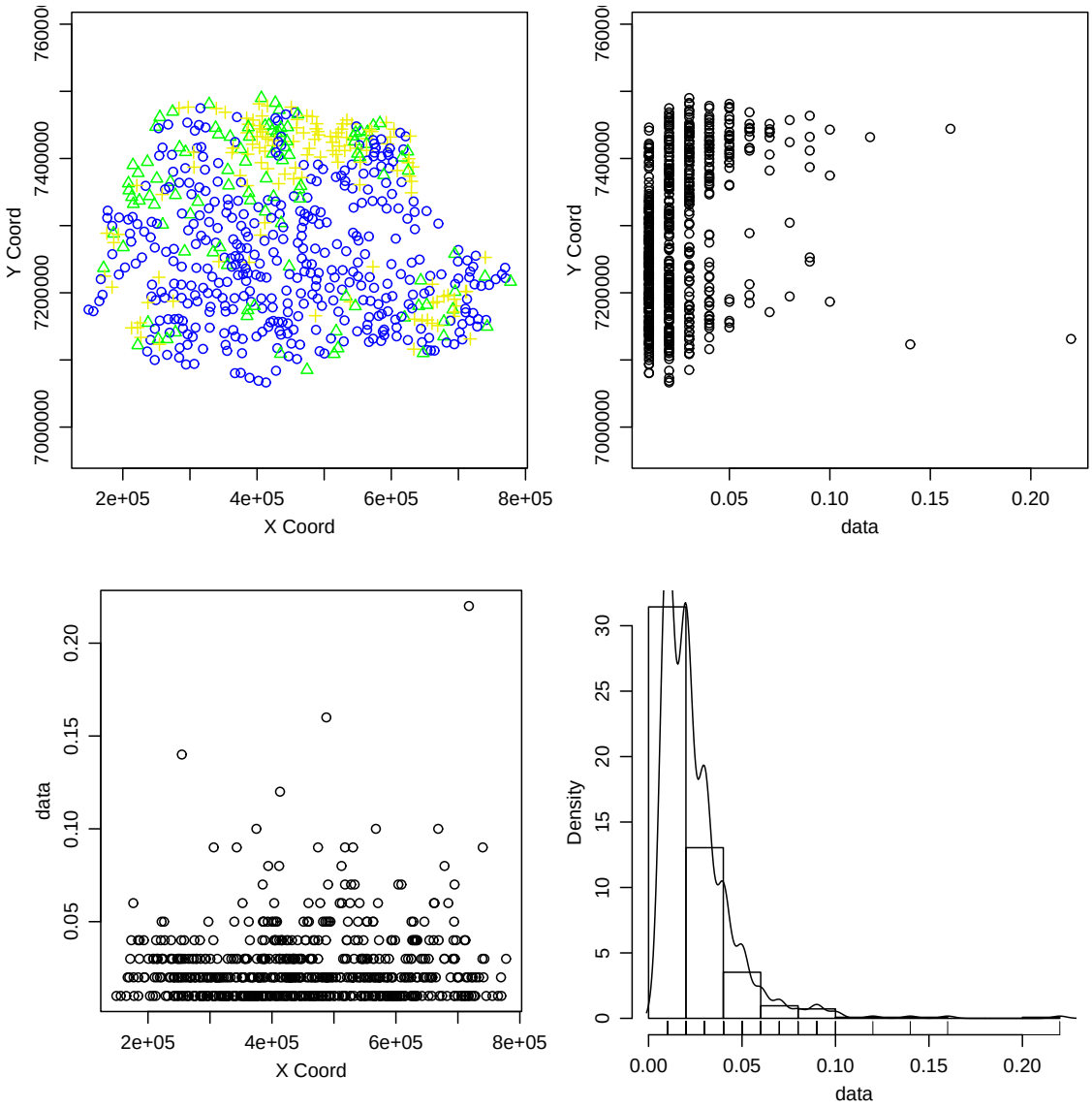

Figure 36: Bromo (Br), removendo dados > 1.

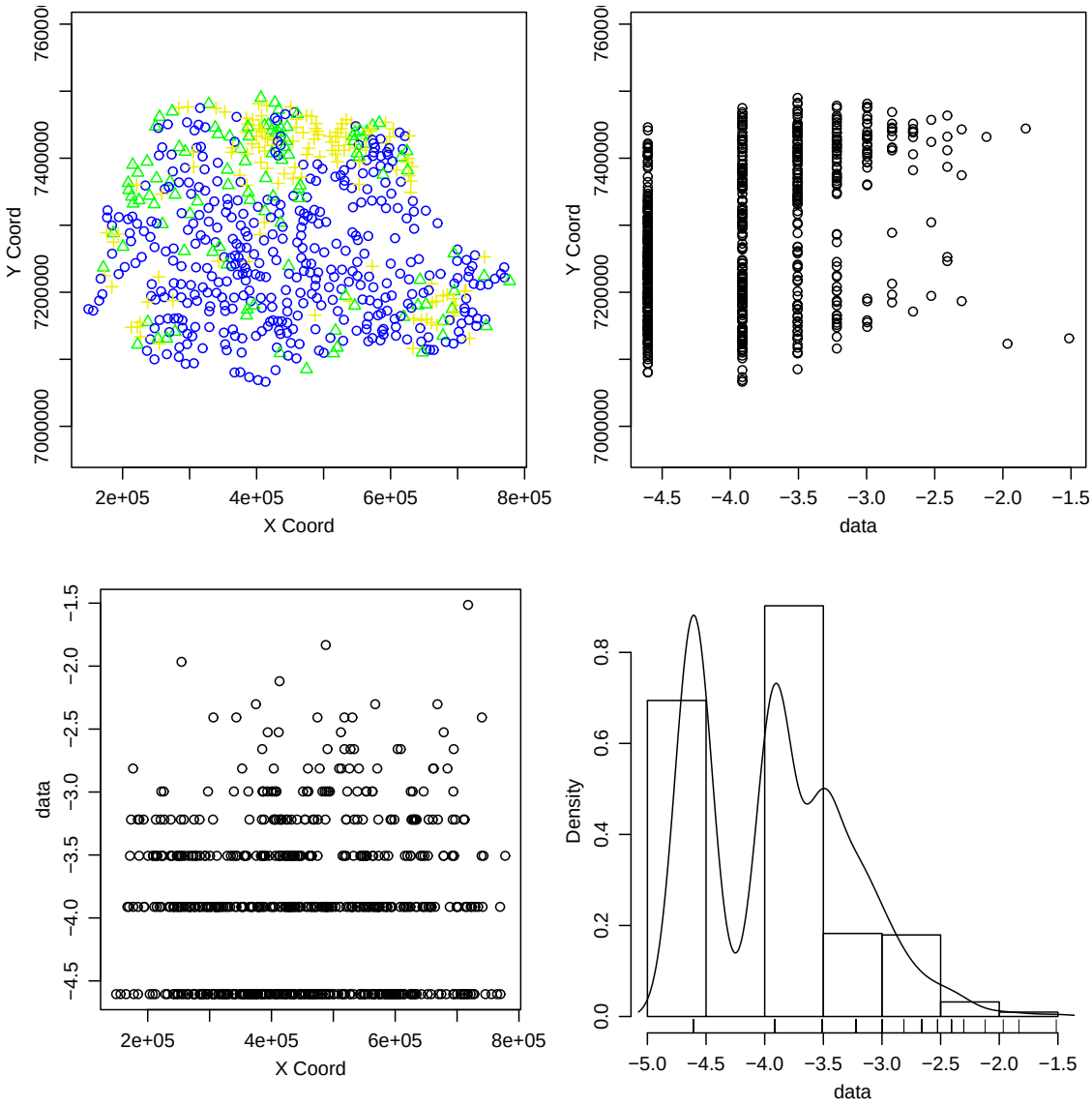

Figure 37: Bromo (Br), retirados dados > 1, transformados (log).

Bromo (Br)

```
as.geodata: 70 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01000 0.01000 0.02000 0.04839 0.03000 8.50000
```

Alumínio (Al)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0200  0.1250  0.1250  0.1733  0.1250  3.8600
```

Bário (Ba)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00700 0.02500 0.02500 0.03494 0.02500 0.27000
```

Indio (In)

```
as.geodata: 83 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05000 0.05000 0.05000 0.06279 0.05000 0.41000
```

Zinco (Zn)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00200 0.01000 0.01000 0.01412 0.01000 2.41000
```

Flúor (F)

```
as.geodata: 2 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.01000 0.02700 0.05257 0.06000 0.98000
```

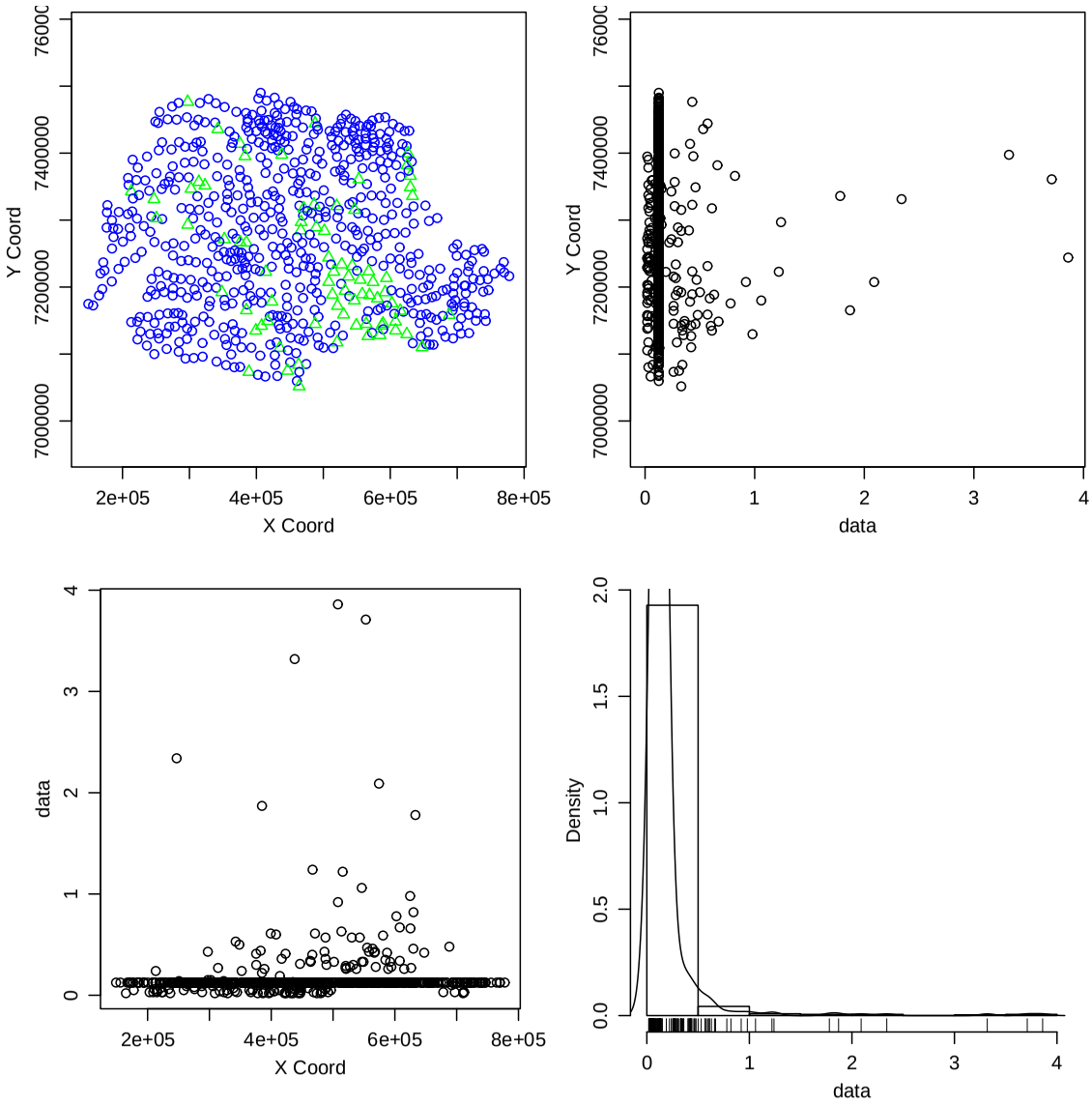

Figure 38: Alumínio (Al), dados originais.

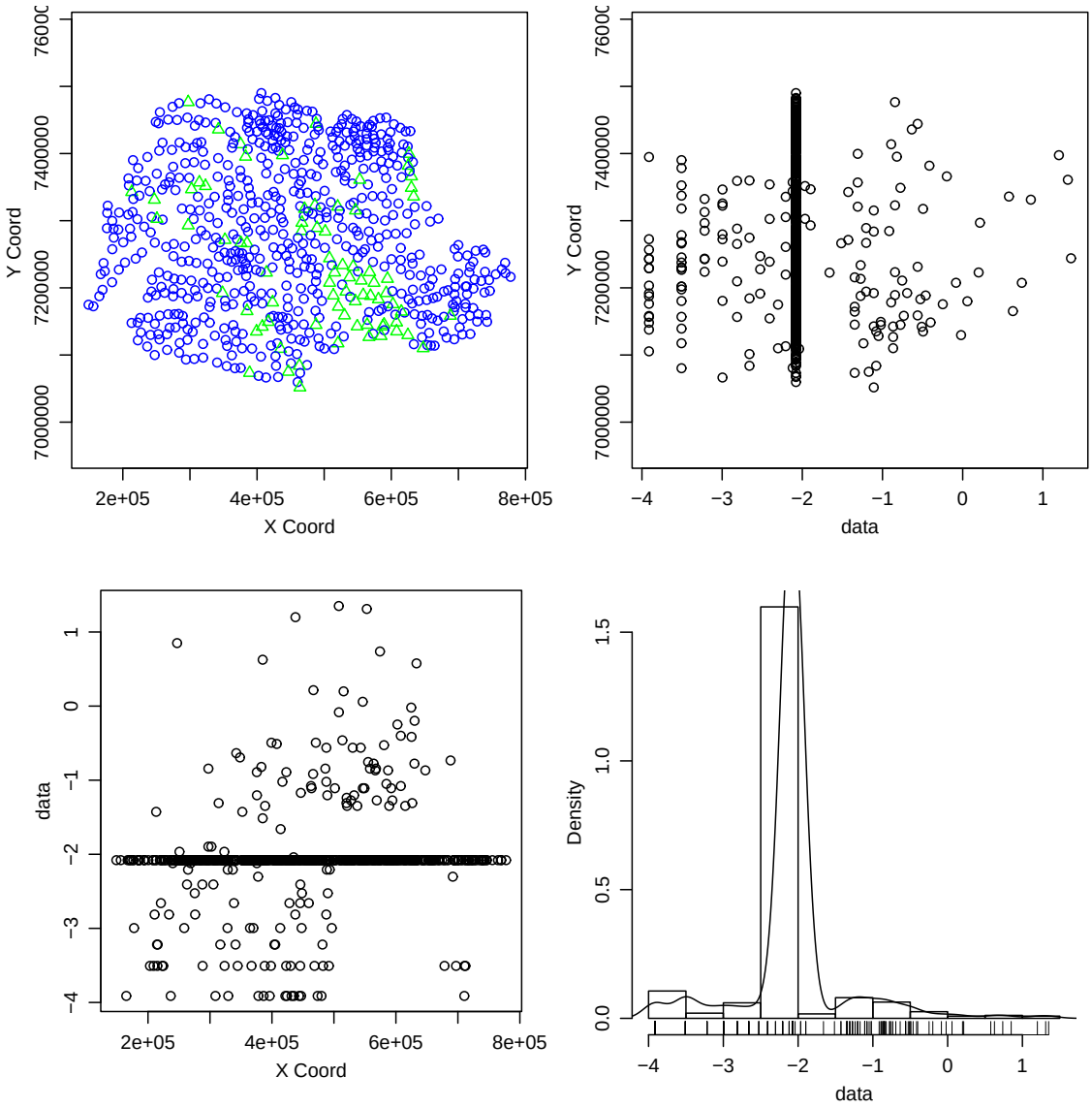

Figure 39: Alumínio (Al), transformados (log).

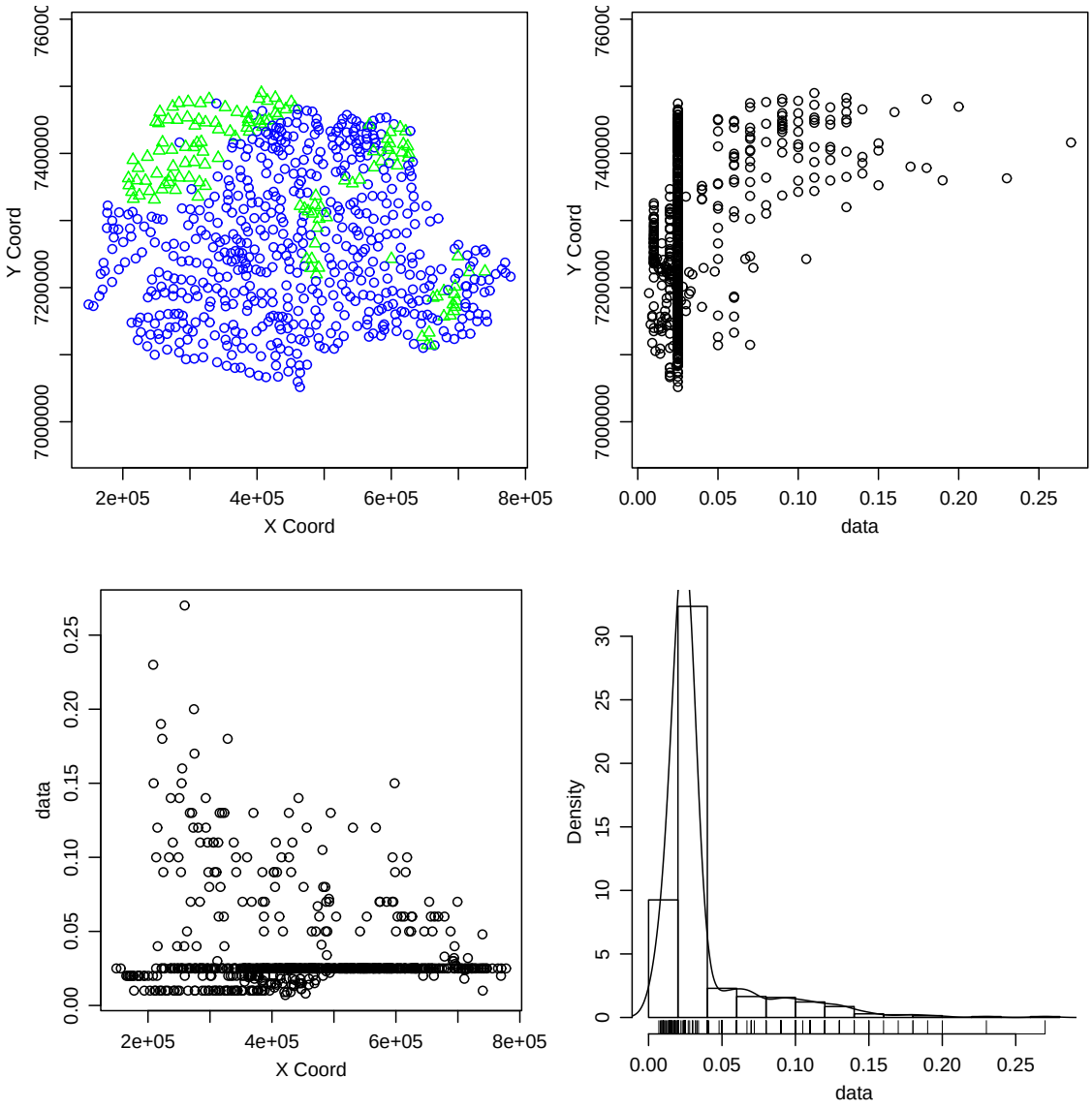

Figure 40: Bário (Ba), dados originais.

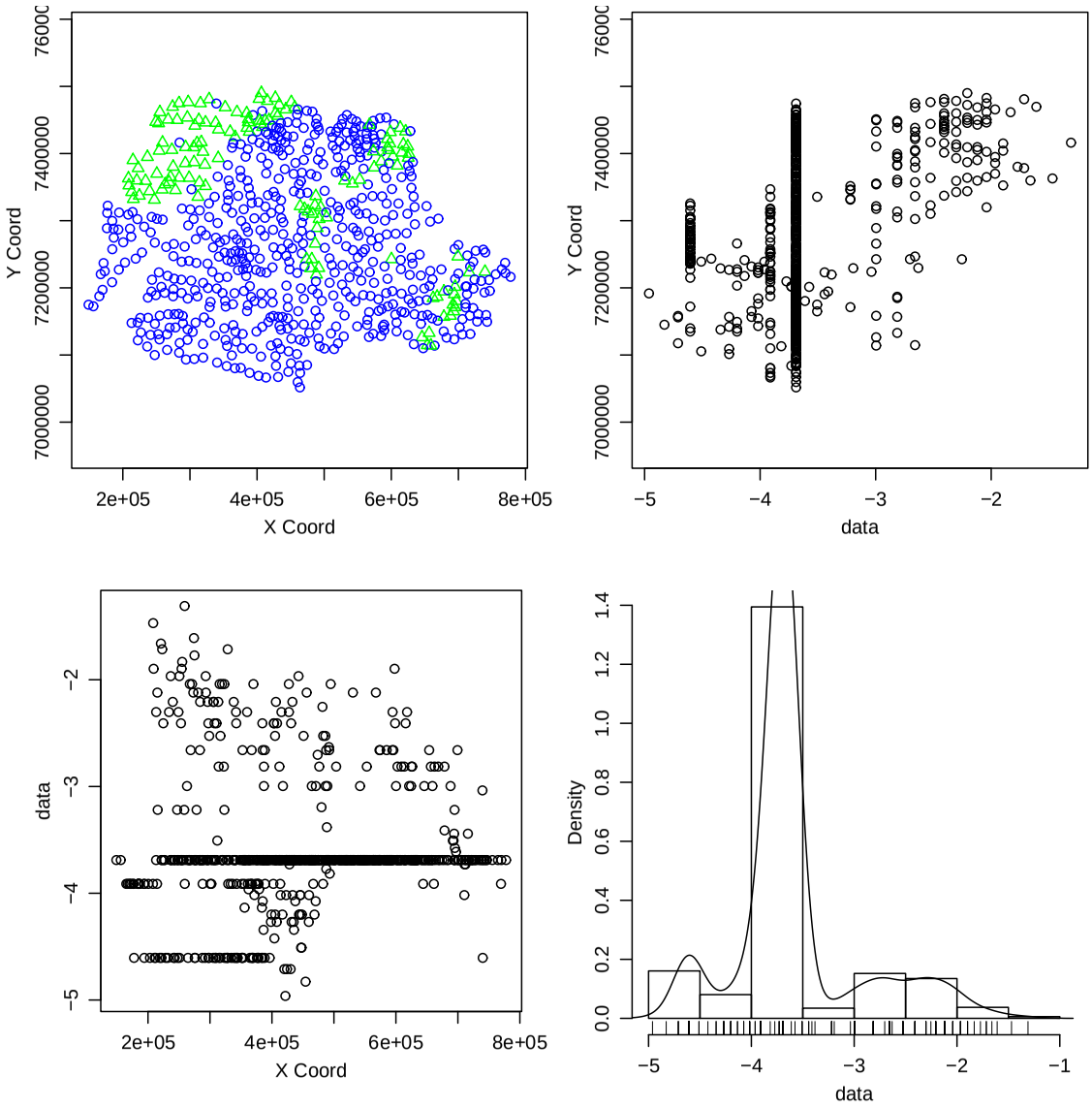

Figure 41: Bário (Ba), dados transformados (log).

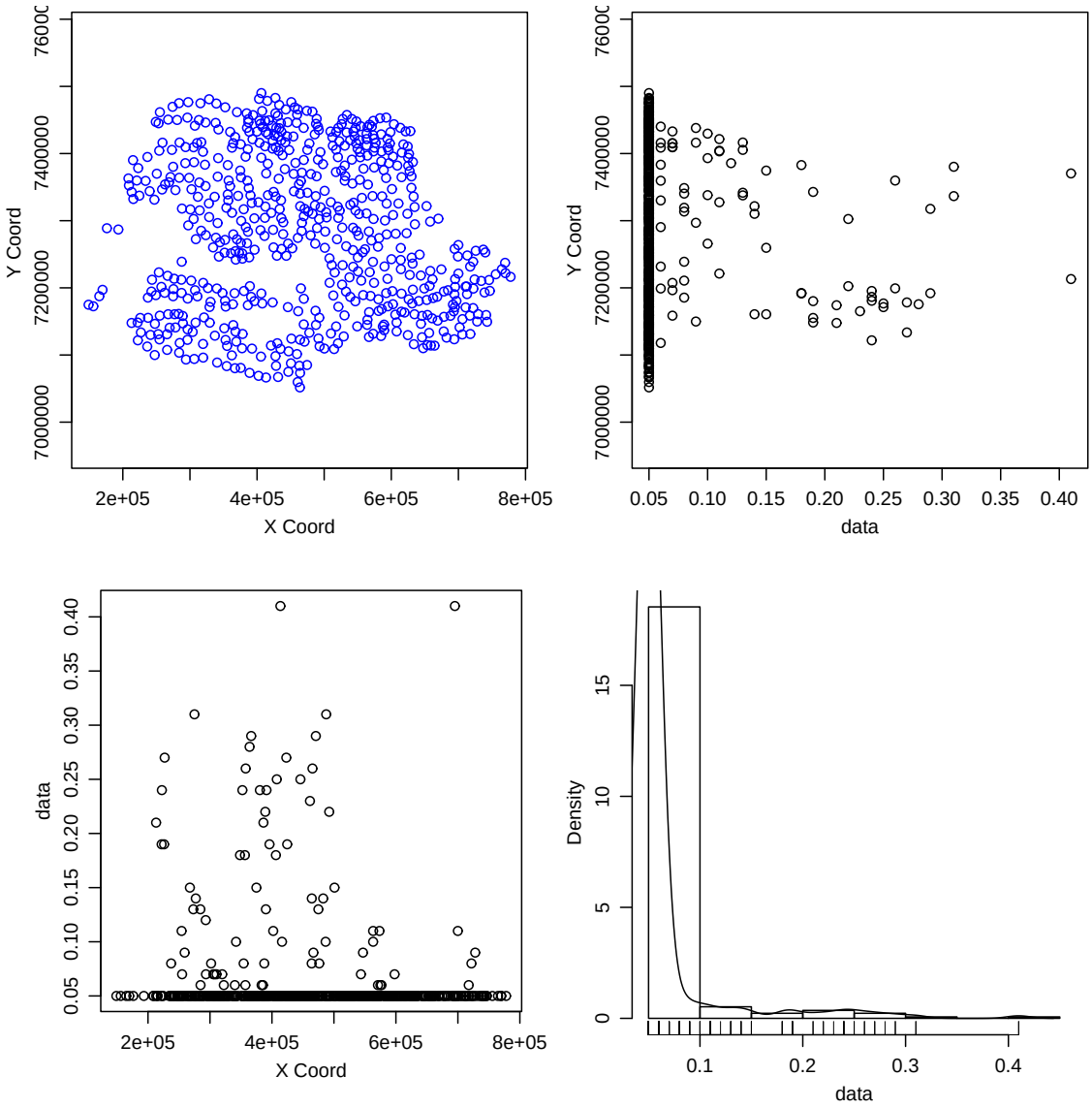

Figure 42: Índio (In), dados originais.

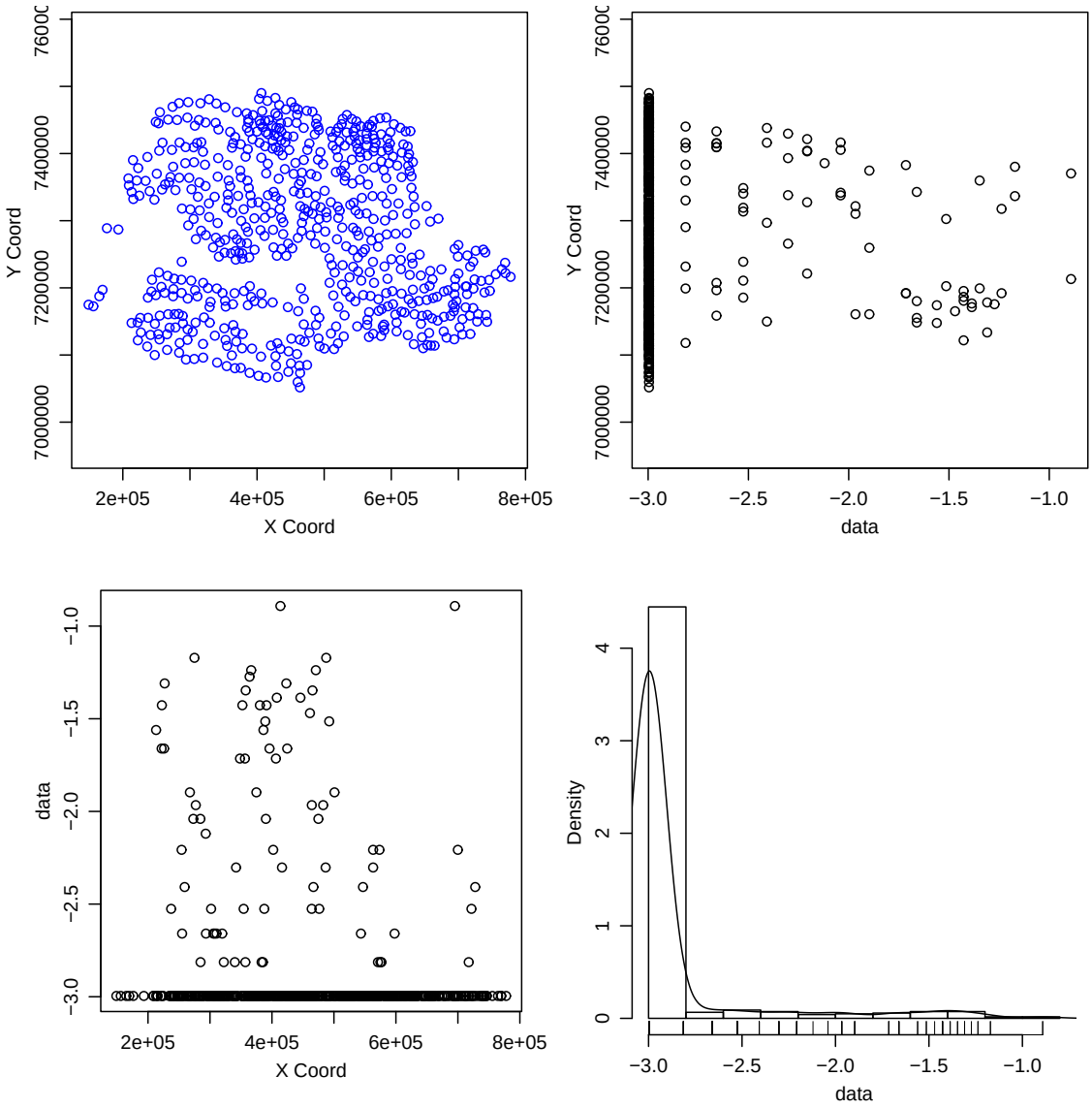

Figure 43: Índio (In), dados transformados (log).

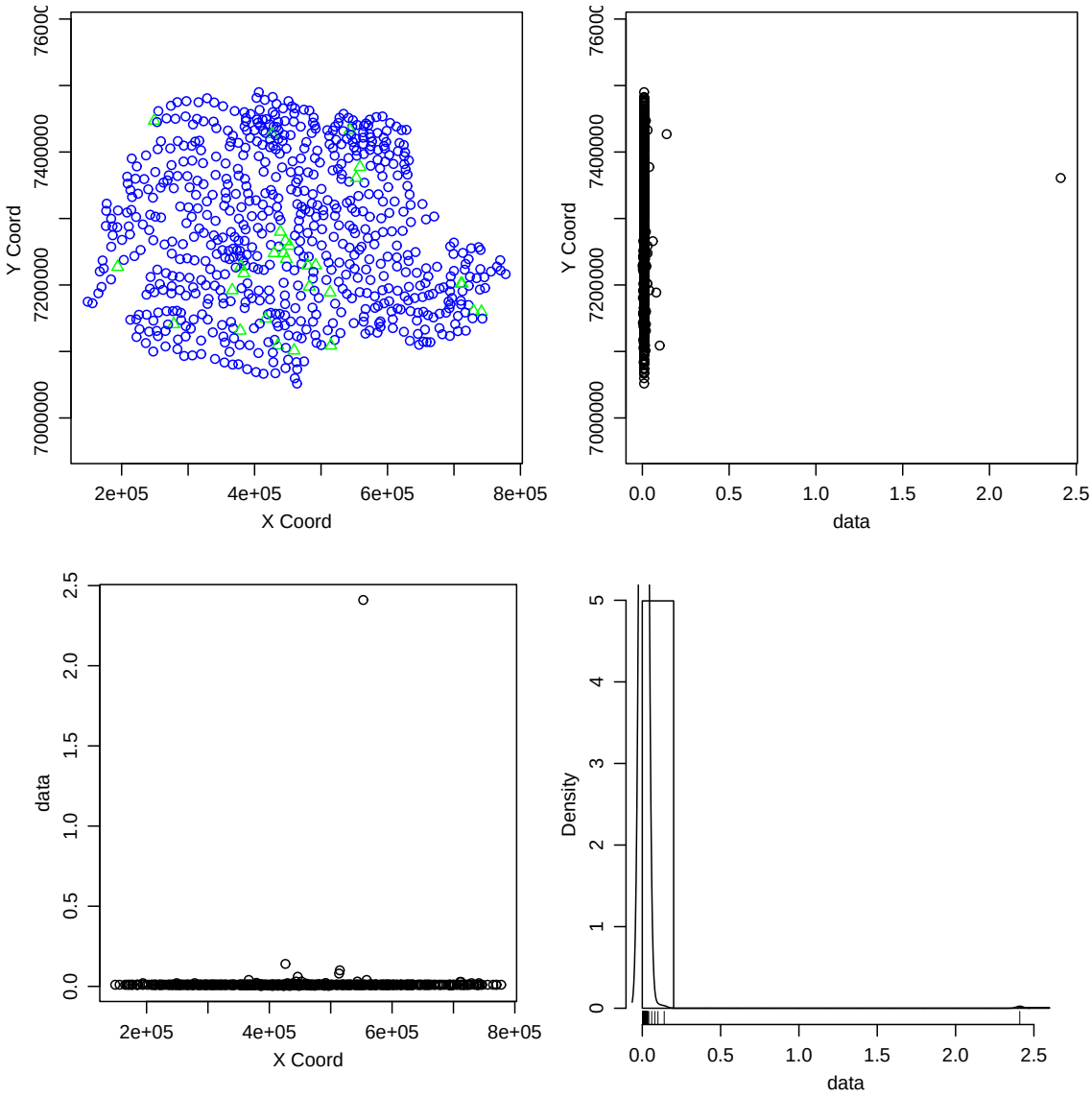

Figure 44: Zinco (Zn), dados originais.

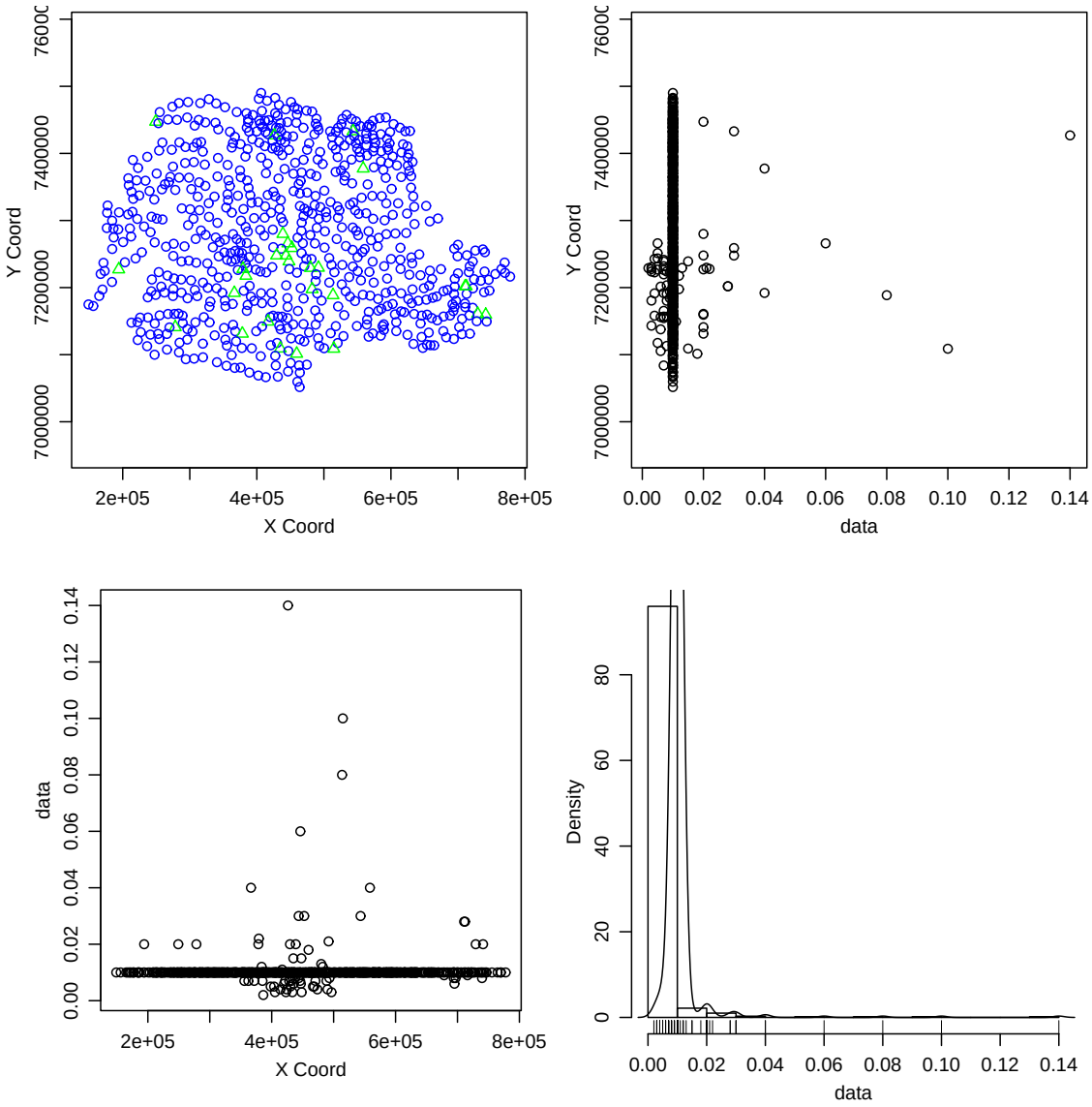

Figure 45: Zinco (Zn), excluindo dados menores que 0,5.

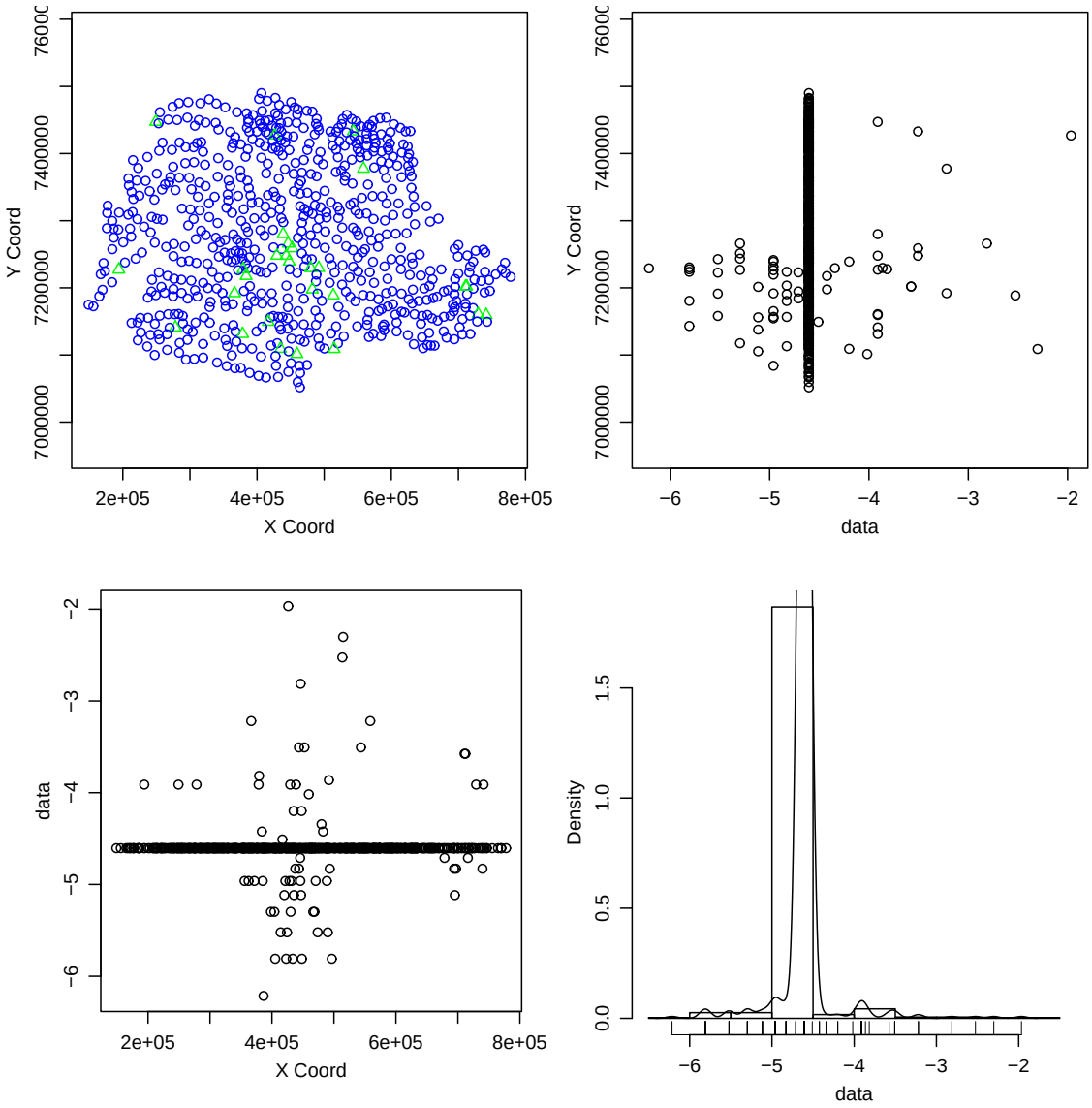

Figure 46: Zinco (Zn), excluindo dados menores que 0,5, transformados (log).

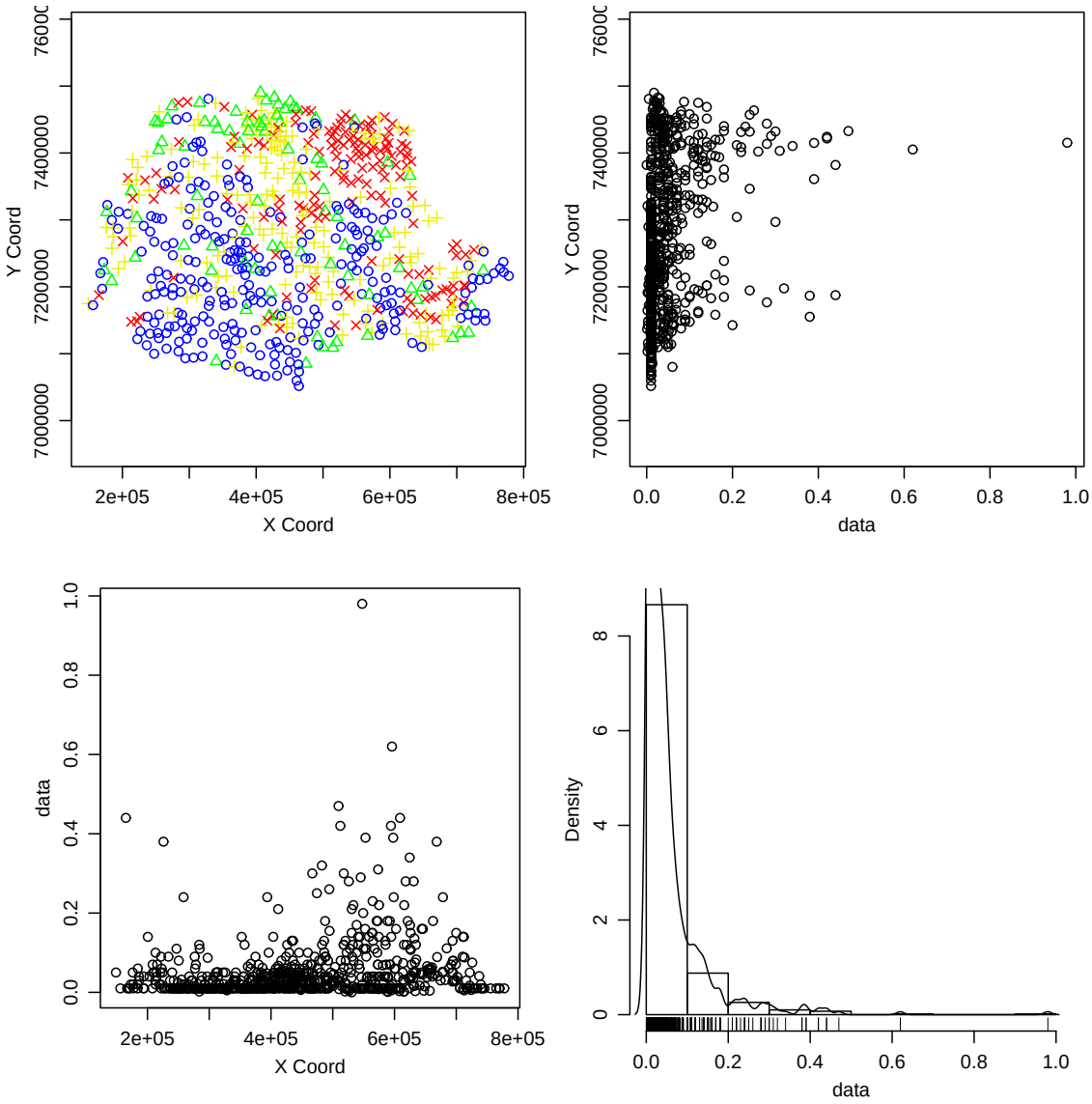

Figure 47: Flúor (F), dados originais.

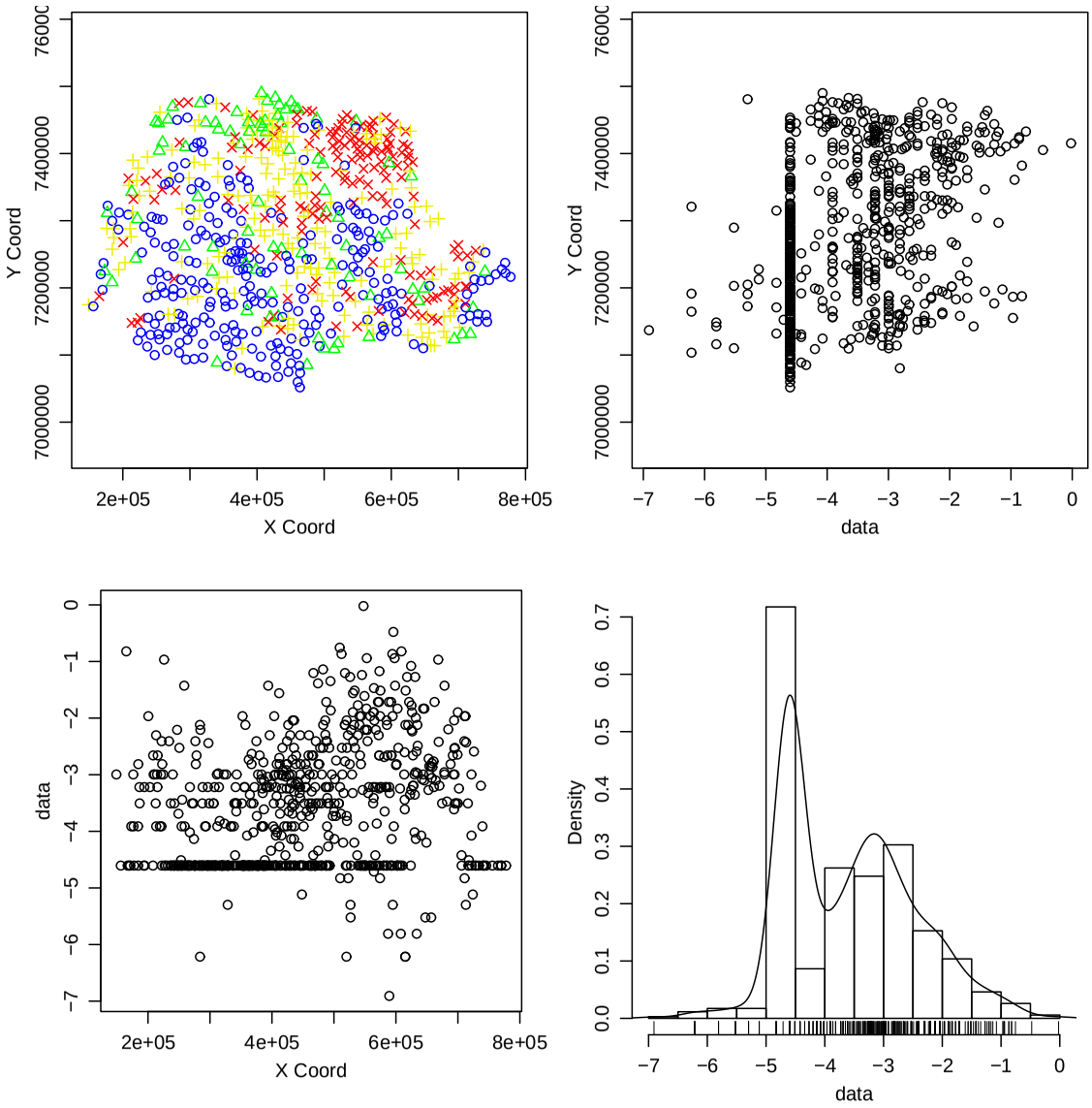

Figure 48: Flúor (F), dados transformados (log).

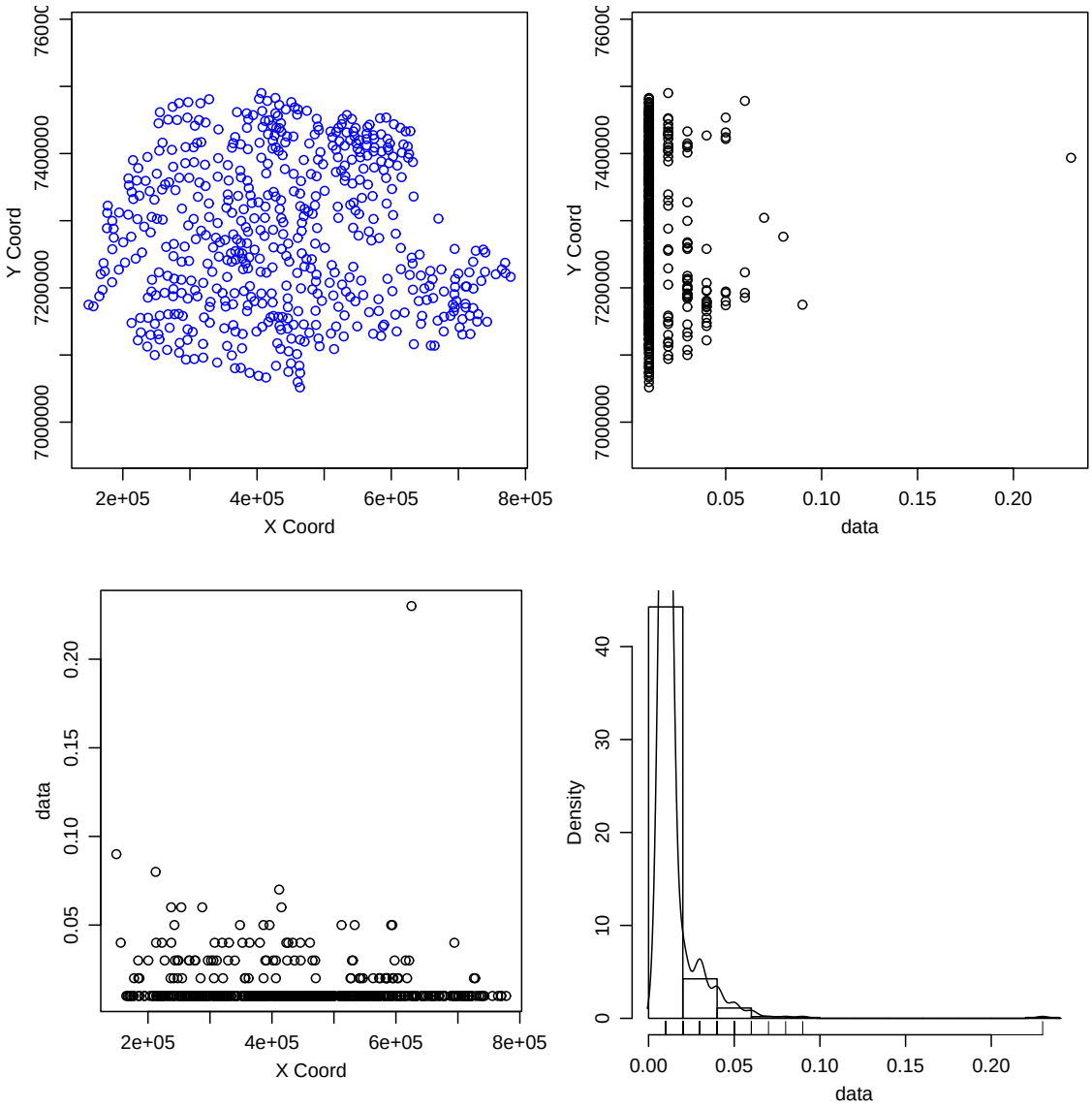

Figure 49: Óxido Nítrico ($N0_2$), dados originais.

Óxido Nítrico (NO_2)

```
as.geodata: 156 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.0100  0.0100  0.0144  0.0100  0.2300
```

1.3 Grupo III

Talvez possam ser, em alguns casos usadas em associação com análise de resíduos dos modelos para taxas de neoplasias.

Prata (Ag)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.025   0.025   0.025   0.025   0.025   0.025
```

Boro (B)

Cádmio (Cd)

Cobalto (Co)

Cromo (Cr)

Cobre (Cu)

Gálio (Ga)

```
as.geodata: 45 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.05    0.05    0.05    0.05    0.05    0.05
```

Molibdênio (Mo)

Níquel (Ni)

Lítio (Li)

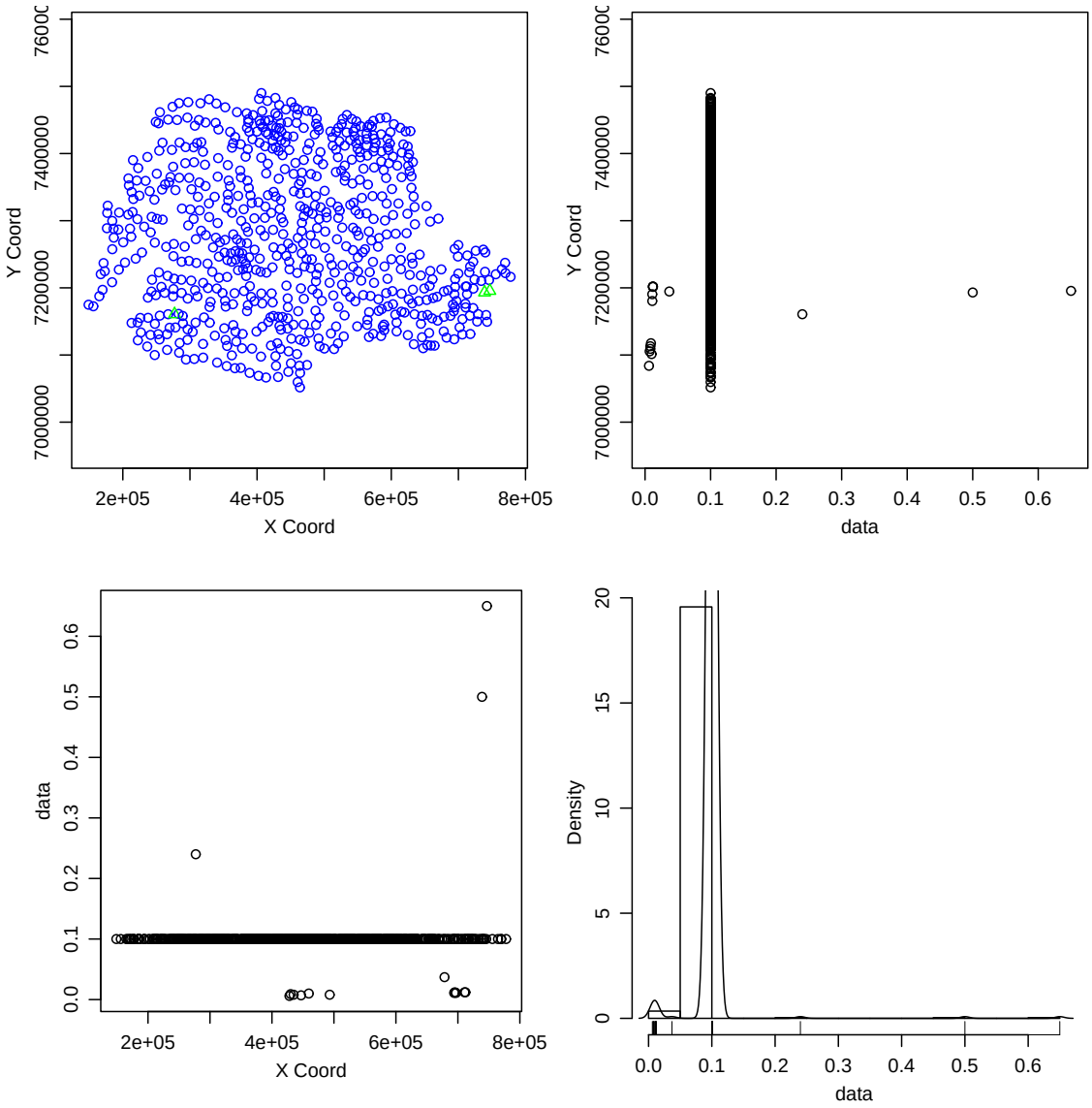

Figure 50: Boro (B), dados originais

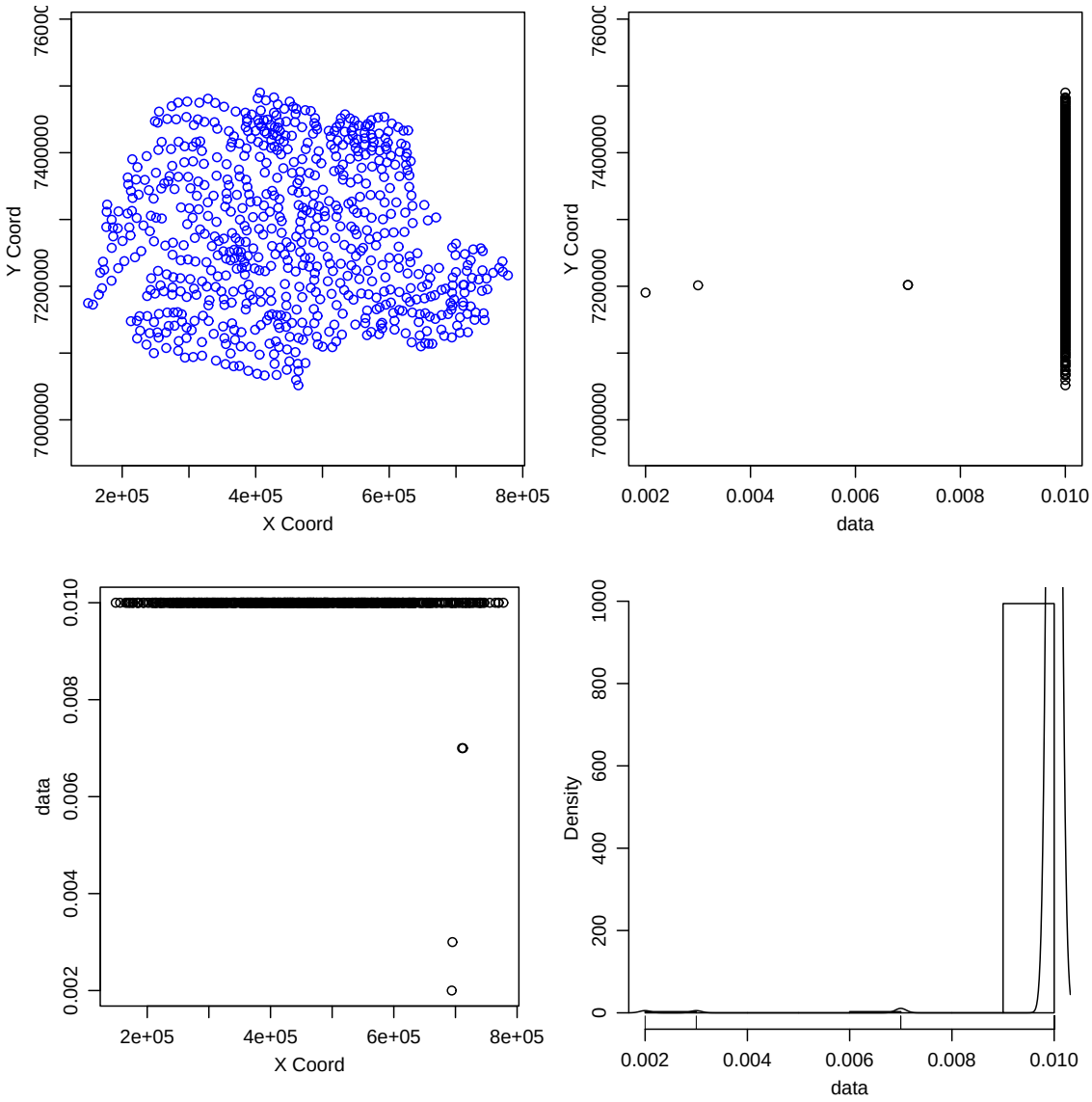

Figure 51: Cádmio (Cd), dados originais

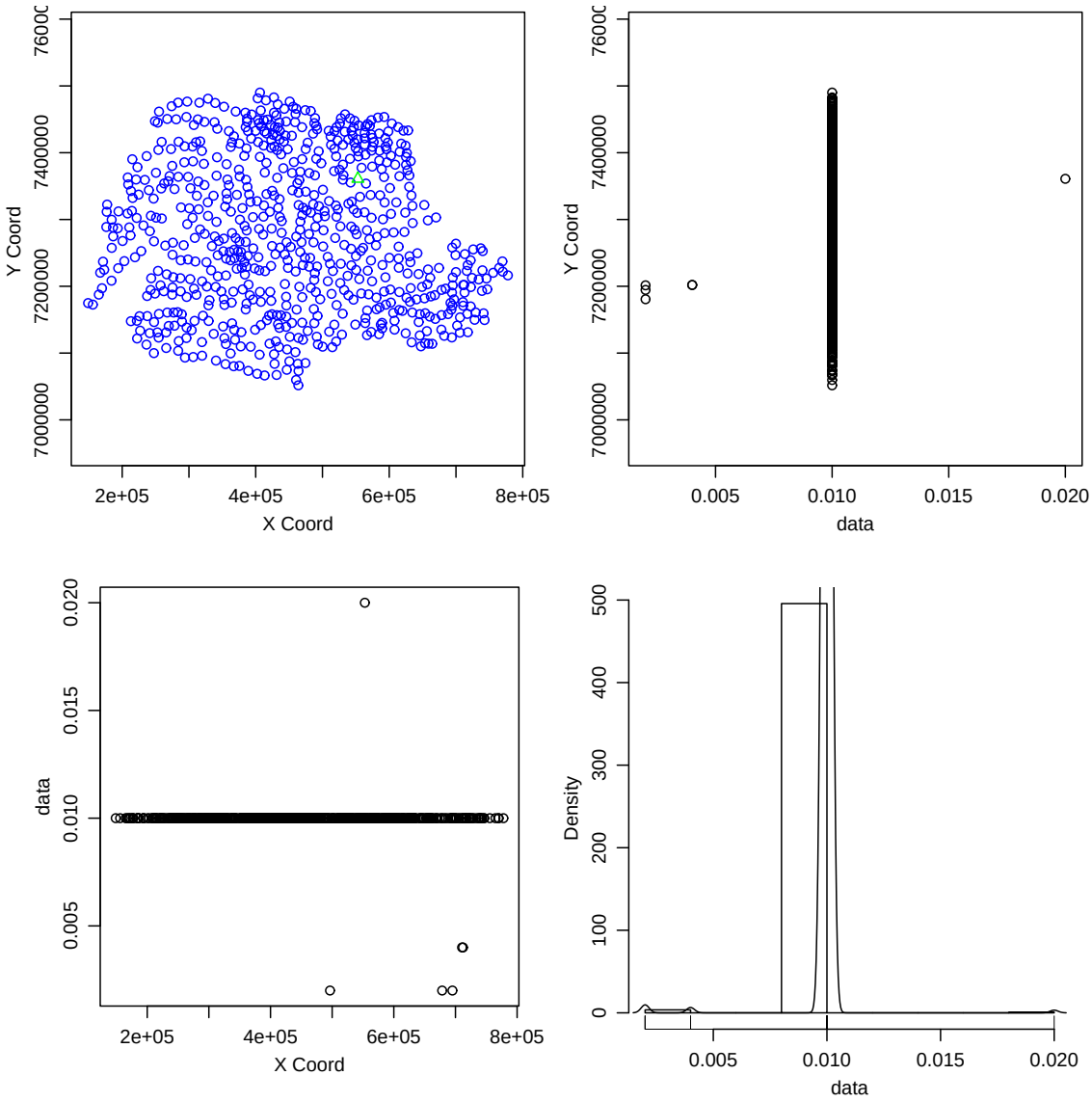

Figure 52: Cobalto (Co), dados originais

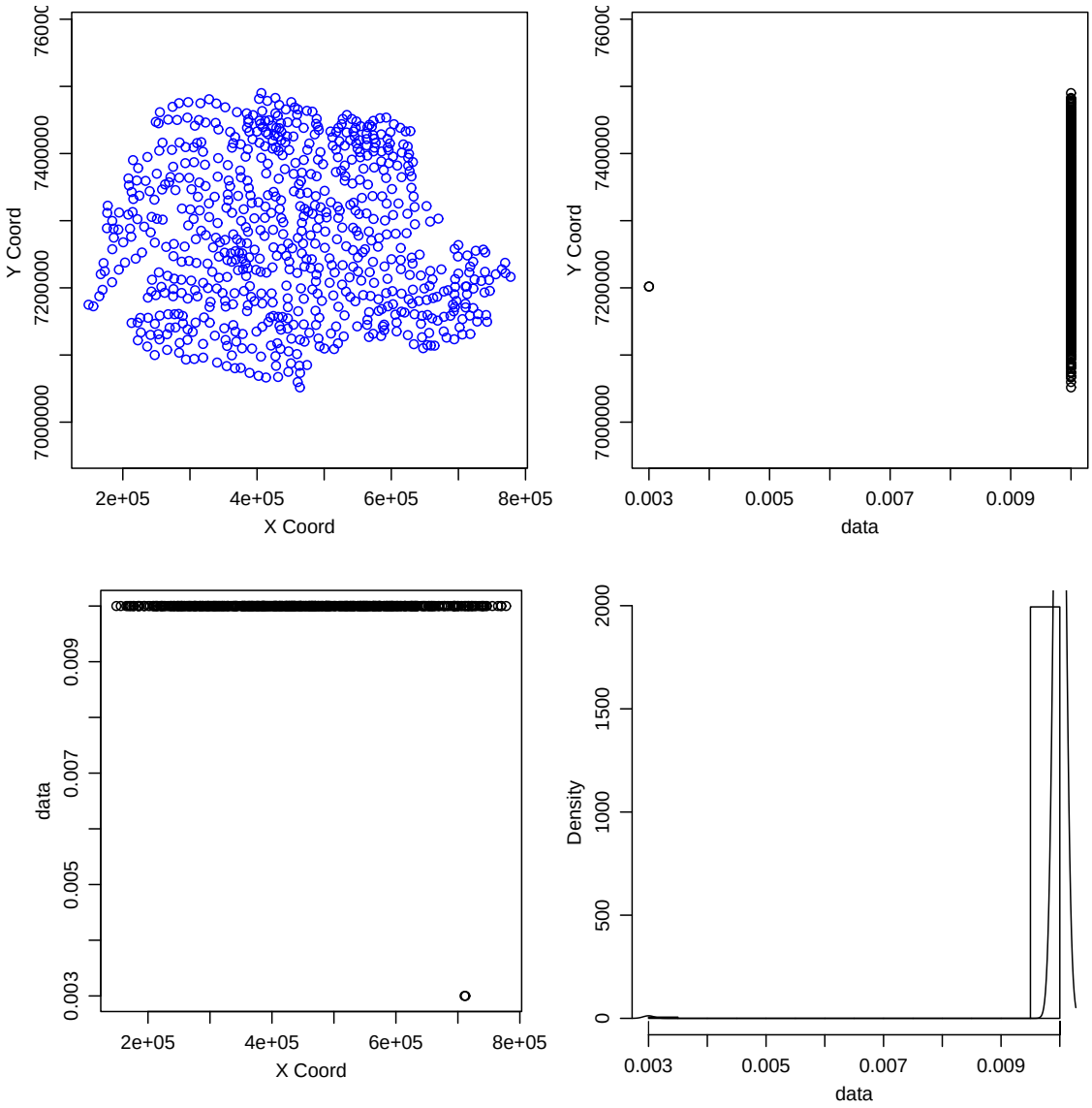

Figure 53: Cromo (Cr), dados originais

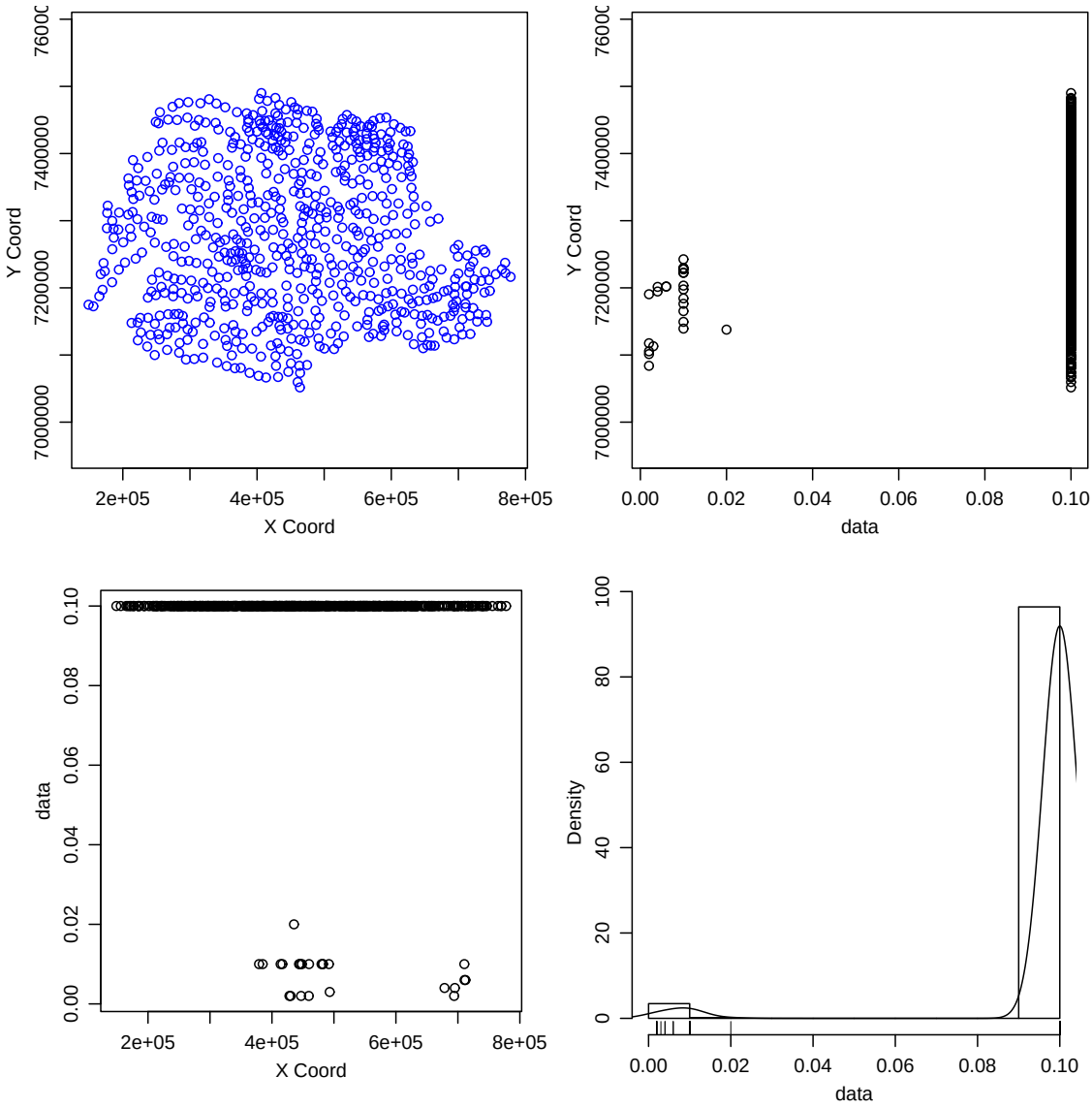

Figure 54: Cobre (Cu), dados originais

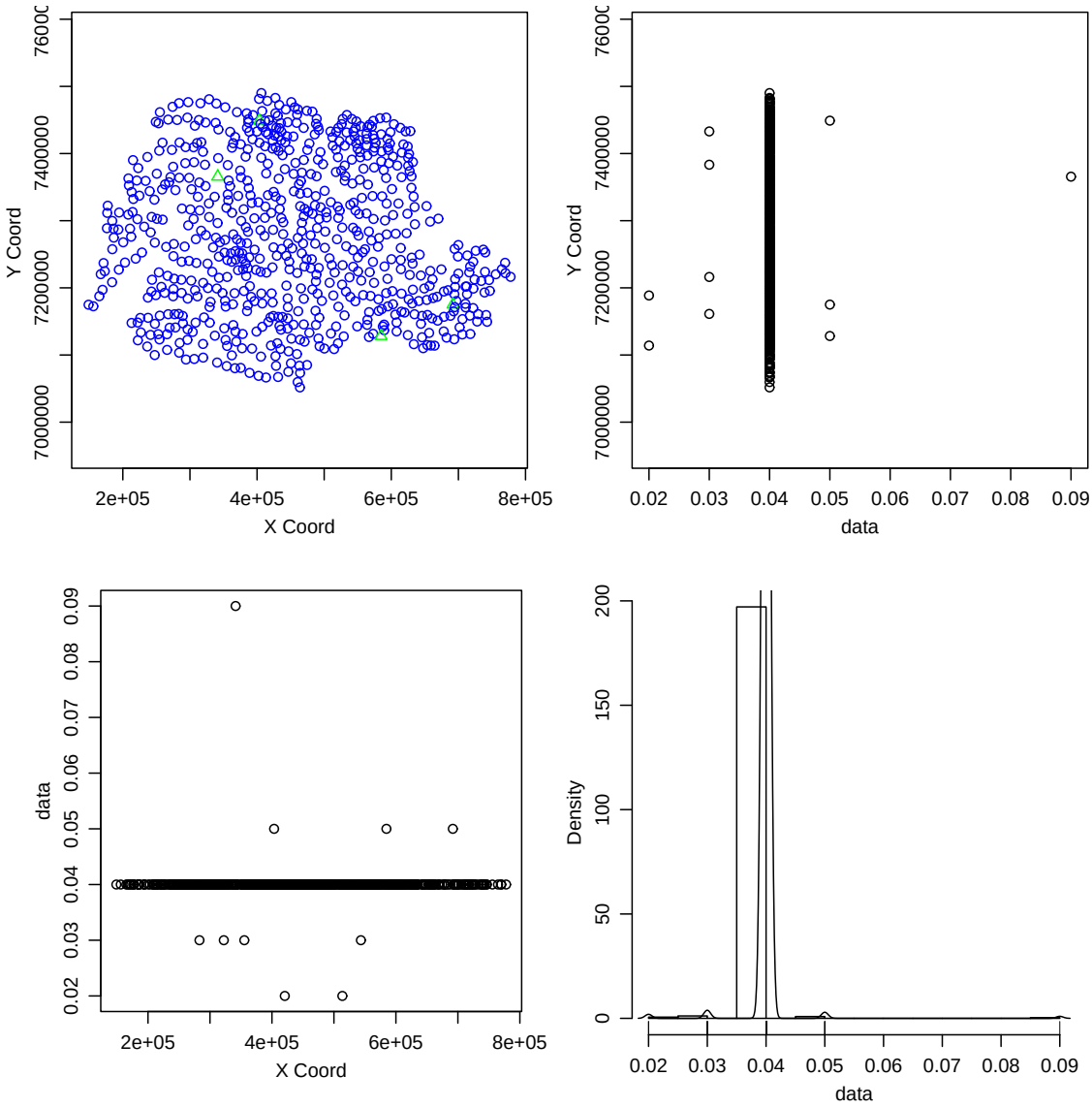

Figure 55: Molibdênio (Mo), dados originais

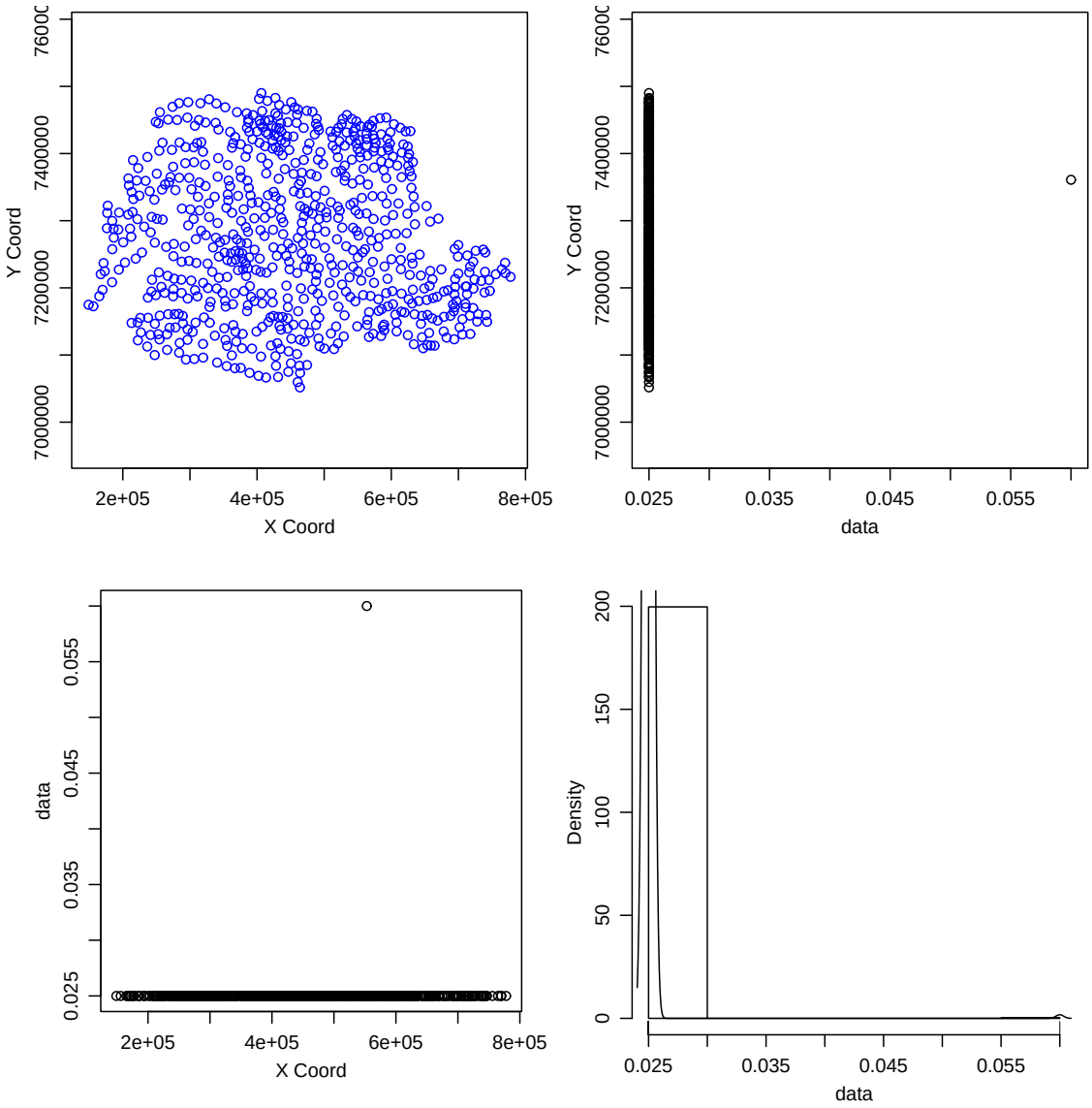

Figure 56: Níquel (Ni), dados originais

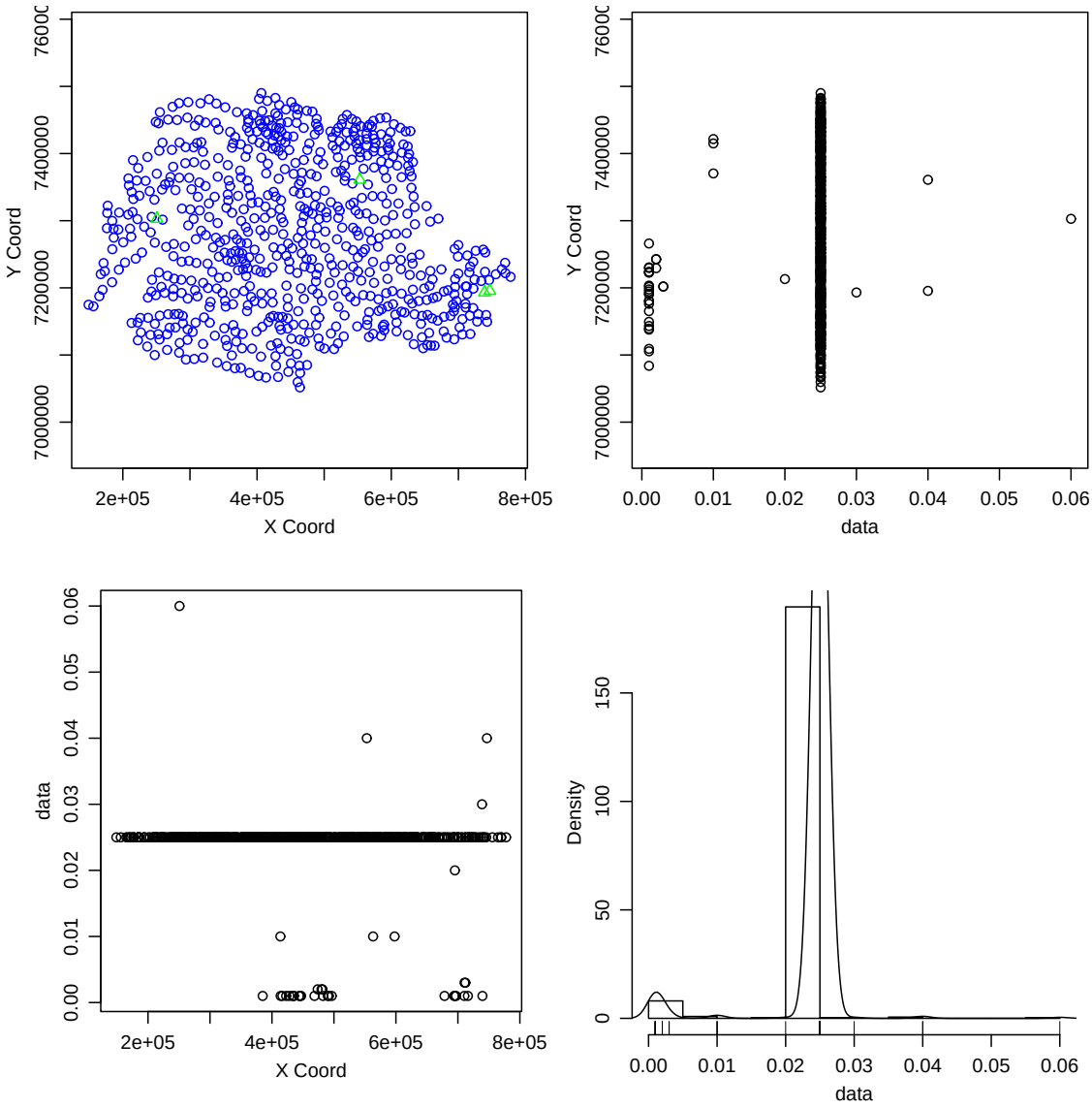

Figure 57: Lítio (Li), dados originais

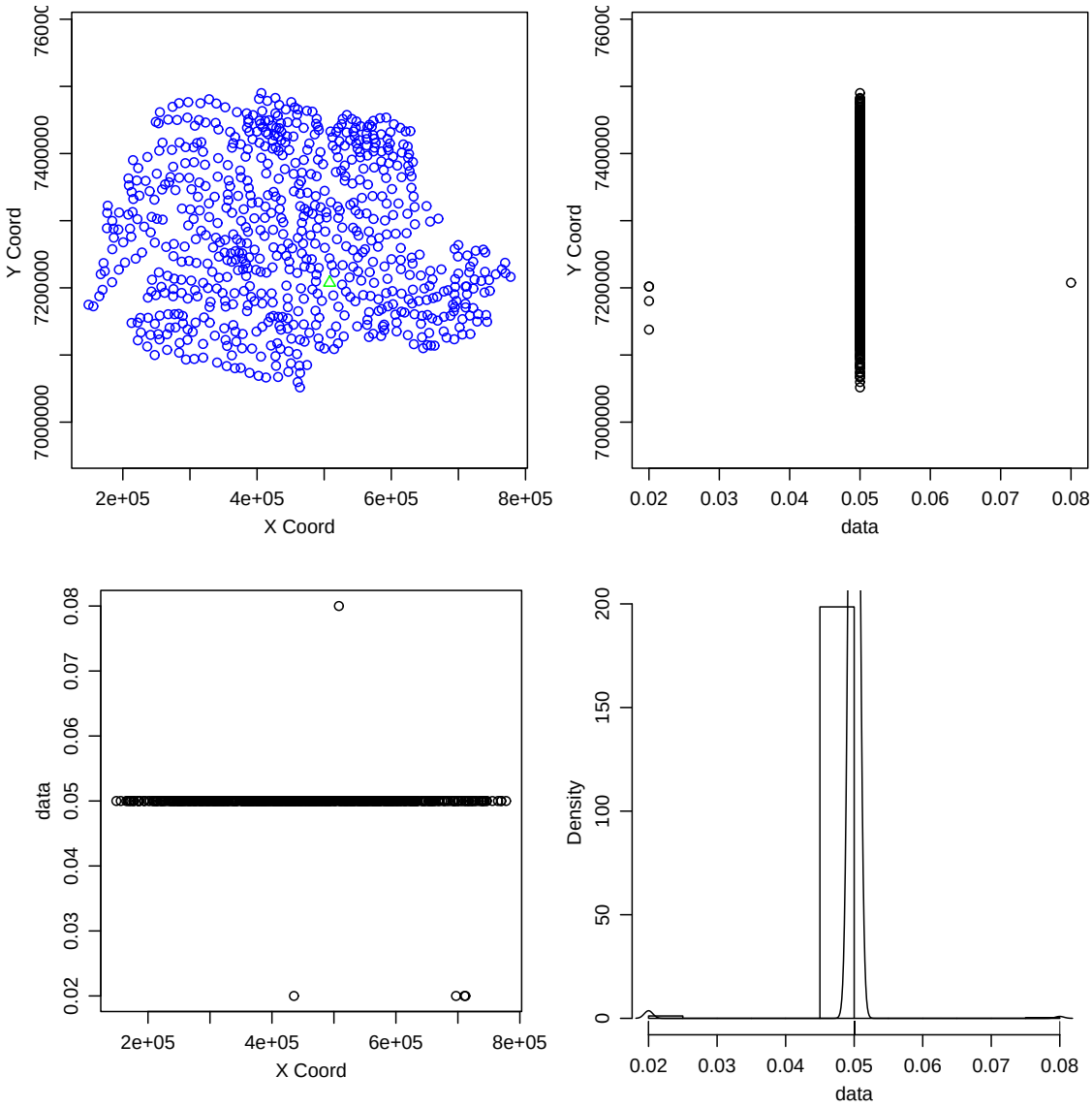

Figure 58: Chumbo (Pb), dados originais

Chumbo (Pb)

Tálio (Tl)

```
as.geodata: 45 points removed due to NA in the data
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.125   0.125   0.125   0.125   0.125   0.125
```

(V)

(W)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.1     0.1     0.1     0.1     0.1     0.1
```

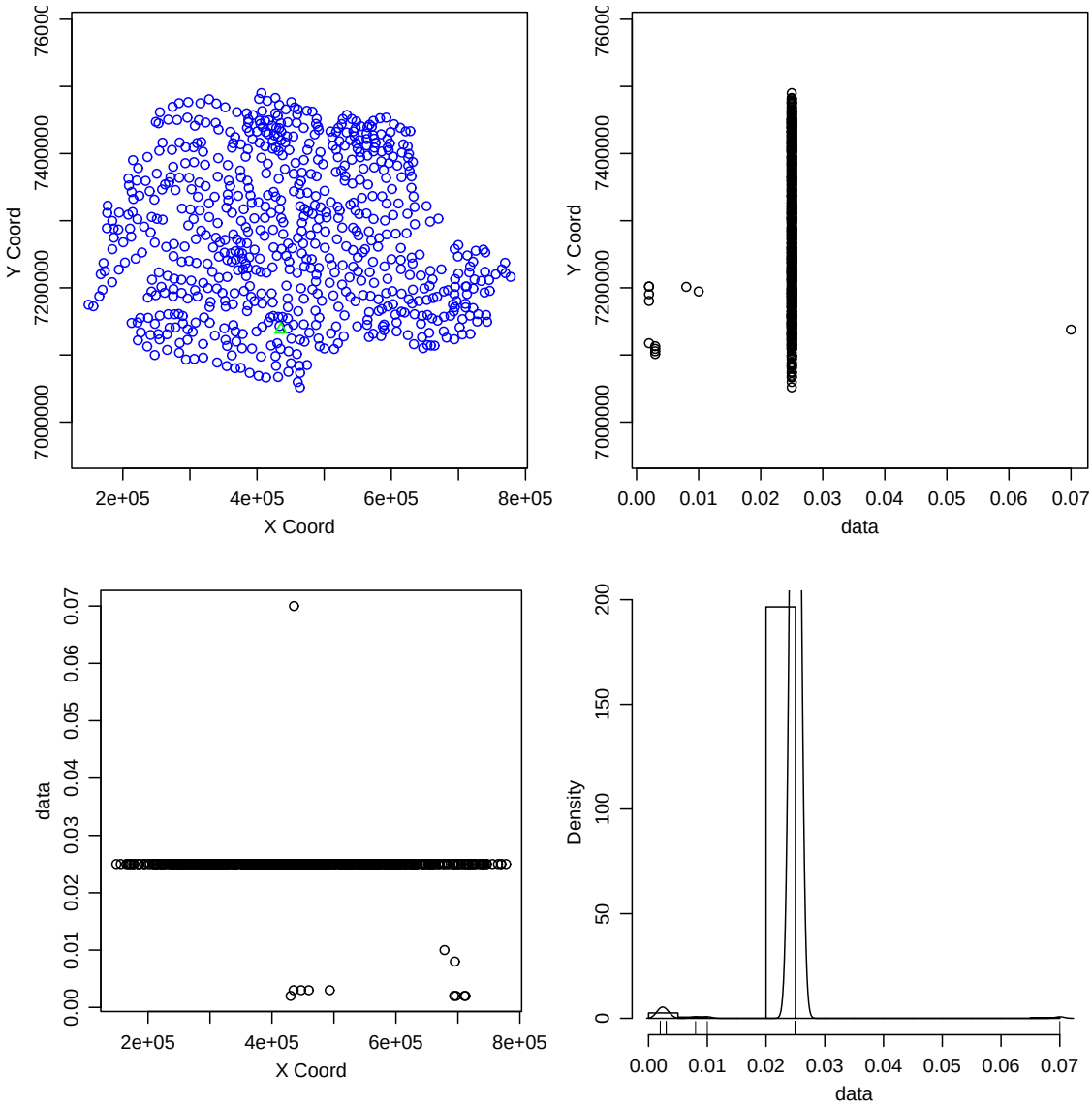

Figure 59: (V), dados originais