

Análise multivariada na caracterização de utilizadores de Internet

M. Rosário de Oliveira

* *Technical University of Lisbon - UTL / Department of Mathematics and CEMAT*

rosario.oliveira@math.ist.utl.pt

**Encontro SPE/CIM com a colaboração do INE em
Probabilidades e Estatística em Telecomunicações**

Coimbra, 19 Maio 2007

Sumário

1. Perfis dos utilizadores da Internet

Sumário

1. Perfis dos utilizadores da Internet

1.1 Descrição do problema

Sumário

1. Perfis dos utilizadores da Internet

1.1 Descrição do problema

1.2 Análise em Componentes Principais

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I
2. Análise da estrutura do tráfego da Internet

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I
2. Análise da estrutura do tráfego da Internet
 - 2.1 Descrição do problema

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I
2. Análise da estrutura do tráfego da Internet
 - 2.1 Descrição do problema
 - 2.2 Classificação dos fluxos próprios

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I
2. Análise da estrutura do tráfego da Internet
 - 2.1 Descrição do problema
 - 2.2 Classificação dos fluxos próprios
 - 2.3 Análise de clusters

Sumário

1. Perfis dos utilizadores da Internet
 - 1.1 Descrição do problema
 - 1.2 Análise em Componentes Principais
 - 1.3 Análise de clusters
 - 1.4 Validação
 - 1.5 Caracterização dos clusters
 - 1.6 Conclusões I
2. Análise da estrutura do tráfego da Internet
 - 2.1 Descrição do problema
 - 2.2 Classificação dos fluxos próprios
 - 2.3 Análise de clusters
 - 2.4 Conclusões II

-
-
-

1. Descrição do problema

Objectivo:

1. Descrição do problema

Objectivo: Encontrar grupos de indivíduos com perfis de utilização da Internet semelhante. O que permite:

1. Descrição do problema

Objectivo: Encontrar grupos de indivíduos com perfis de utilização da Internet semelhante. O que permite:

- Fornecer serviços diferenciados de acordo com o comportamento horário dos utilizadores.

1. Descrição do problema

Objectivo: Encontrar grupos de indivíduos com perfis de utilização da Internet semelhante. O que permite:

- Fornecer serviços diferenciados de acordo com o comportamento horário dos utilizadores.
- Optimizar recursos.

1. Descrição do problema

Objectivo: Encontrar grupos de indivíduos com perfis de utilização da Internet semelhante. O que permite:

- Fornecer serviços diferenciados de acordo com o comportamento horário dos utilizadores.
- Optimizar recursos.
- Propor sistemas tarifários que promovam o uso do Internet a horas de menos tráfego.

1. Descrição do problema (cont.)

Analisaram-se dados recolhidos em dois ISPs portugueses distintos.

1. Descrição do problema (cont.)

Analisaram-se dados recolhidos em dois ISPs portugueses distintos.

Ambos os ISPs oferecem diferentes tipos de contratos, caracterizados por diferentes valores máximos da taxa de transferência no sentido downstream/upstream (em Kbits/s)

1. Descrição do problema (cont.)

Analisaram-se dados recolhidos em dois ISPs portugueses distintos.

Ambos os ISPs oferecem diferentes tipos de contratos, caracterizados por diferentes valores máximos da taxa de transferência no sentido downstream/upstream (em Kbits/s)

— ISP1: 128/64, 256/128 e 512/256.

1. Descrição do problema (cont.)

Analisaram-se dados recolhidos em dois ISPs portugueses distintos.

Ambos os ISPs oferecem diferentes tipos de contratos, caracterizados por diferentes valores máximos da taxa de transferência no sentido downstream/upstream (em Kbits/s)

- ISP1: 128/64, 256/128 e 512/256.
- ISP2: 512/128 e 1024/256.

-
-
-

1. Descrição do problema (cont.)

-
-
-

1. Descrição do problema (cont.)

Variáveis:

1. Descrição do problema (cont.)

Variáveis: Taxa de transferência em Kbit/s, no sentido *downstream*, calculada em intervalos de 30 minutos, durante um Sábado.

1. Descrição do problema (cont.)

Variáveis: Taxa de transferência em Kbit/s, no sentido *downstream*, calculada em intervalos de 30 minutos, durante um Sábado.

$$(X_1, \dots, X_{48}),$$

1. Descrição do problema (cont.)

Variáveis: Taxa de transferência em Kbit/s, no sentido *downstream*, calculada em intervalos de 30 minutos, durante um Sábado.

$$(X_1, \dots, X_{48}),$$

onde X_i representa taxa de transferência na i -ésima meia hora do Sábado em estudo.

-
-
-

1. Descrição do problema (cont.)

Observações:

1. Descrição do problema (cont.)

Observações: Dois conjuntos de dados

1. Descrição do problema (cont.)

Observações: Dois conjuntos de dados

ISP1 (CATV) and **ISP2** (ADSL)

1. Descrição do problema (cont.)

Observações: Dois conjuntos de dados

ISP1 (CATV) and **ISP2** (ADSL)

	ISP1	ISP2
Data captura	Nov. 9, 2002	Out. 19, 2002
Nº clientes	3432	875
Downloaded MBytes por cliente	85.6	489
Nº fluxos por utilizador	1.80	1.75
Duração fluxos (horas)	4.9	8.6
Taxa média de débito (Kbits/sec)	23.9	81.1

1. Descrição do problema (cont.)

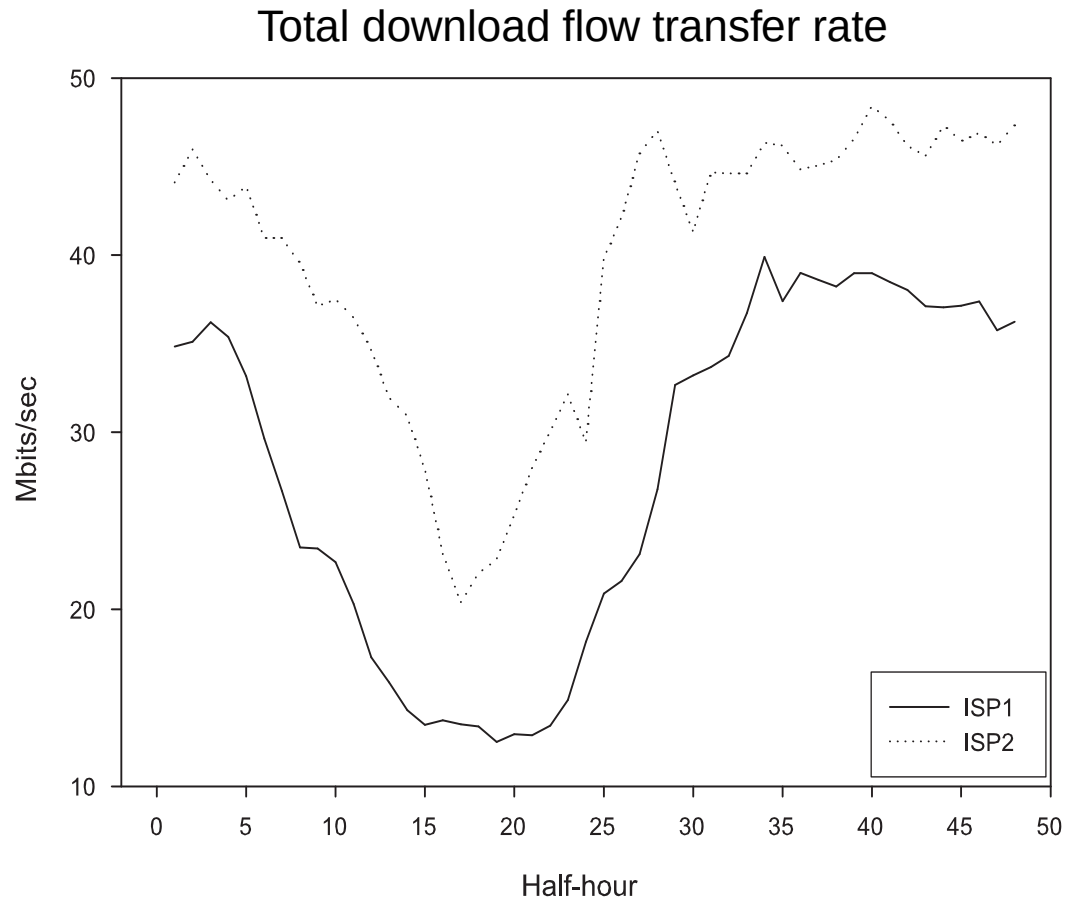
As aplicações identificadas foram agrupadas do seguinte modo:

1. Descrição do problema (cont.)

As aplicações identificadas foram agrupadas do seguinte modo:

- **File sharing**: Kazaa (TCP 1214), eDonkey2000 (TCP 4662), direct connect (TCP 412, 1412), WinMX (TCP 6699) e FTP (TCP 20, 21).
- **HTTP**: HTTP (80) e secure HTTP (443).
- **Games**: Operation FlashPoint (TCP 2234), Medal of Honor (UDP 12203), Half-Life (UDP 27005) e todo o tráfego associado a Game Connection Port (TCP 2346).
- **IRC/news**: IRC traffic (TCP 1025, 1026) e Newsgroups access (TCP 119).
- **Mail**: POP3 mail access (TCP 110).
- **Streaming**: MS-Streaming (TCP 1755) e RealPlayer (UDP 6970).
- **Others**: outras aplicações e tráfego não identificado.

1. Descrição do problema (cont.)



2. Análise em componentes principais

Os dados sugerem que à medida que a taxa média de transferência diária aumenta a variabilidade entre o que acontece de manhã e à tarde também aumenta.

2. Análise em componentes principais

Os dados sugerem que à medida que a taxa média de transferência diária aumenta a variabilidade entre o que acontece de manhã e à tarde também aumenta.

Para estabilizar esta variabilidade considera-se a transformação:

2. Análise em componentes principais

Os dados sugerem que à medida que a taxa média de transferência diária aumenta a variabilidade entre o que acontece de manhã e à tarde também aumenta.

Para estabilizar esta variabilidade considera-se a transformação:

$$Y_i = \ln(1 + X_i)$$

2. Análise em componentes principais

Com o intuito de visualizar os dados construíram-se componentes principais,

2. Análise em componentes principais

Com o intuito de visualizar os dados construíram-se componentes principais,

$$CP_i = \alpha_{i1}Y_1 + \dots + \alpha_{i,48}Y_{48},$$

2. Análise em componentes principais

Com o intuito de visualizar os dados construíram-se componentes principais,

$$CP_i = \alpha_{i1}Y_1 + \dots + \alpha_{i,48}Y_{48},$$

i.e., projectaram-se as observações nas duas dimensões, ortogonais, que maximizam a variabilidade observada.

2. Análise em componentes principais

Com o intuito de visualizar os dados construíram-se componentes principais,

$$CP_i = \alpha_{i1}Y_1 + \dots + \alpha_{i,48}Y_{48},$$

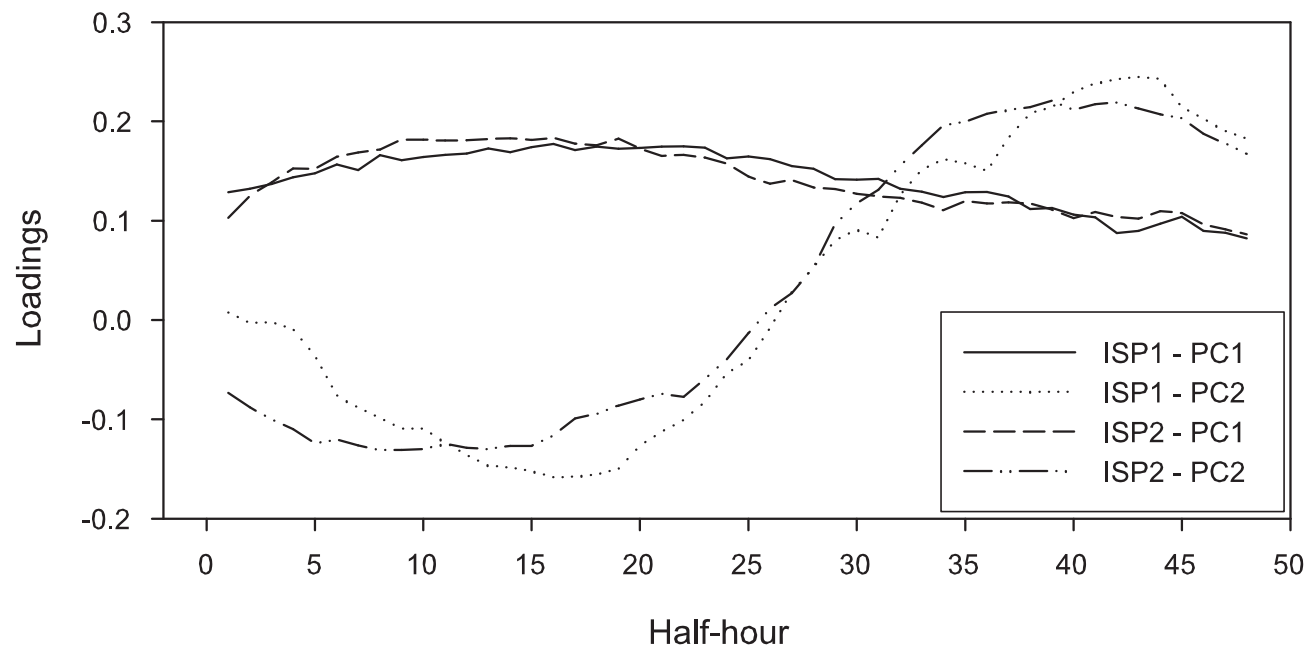
i.e., projectaram-se as observações nas duas dimensões, ortogonais, que maximizam a variabilidade observada.

As duas primeiras componentes principais explicam:

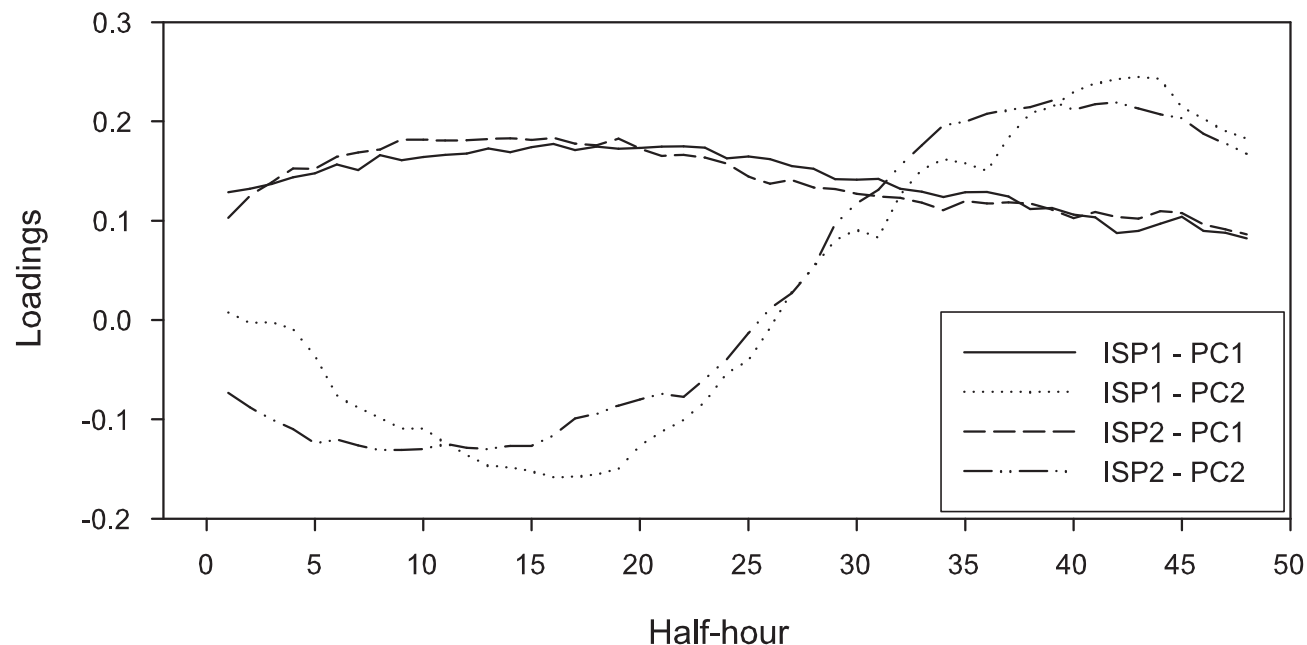
ISP1 — 50.6% da variabilidade total do ISP1

ISP2 — 60.4% da variabilidade total do ISP2

2. Análise em componentes principais (cont.)



2. Análise em componentes principais (cont.)



Interpretação:

CP_1 – Média ponderada das taxas de transf. ao longo do d

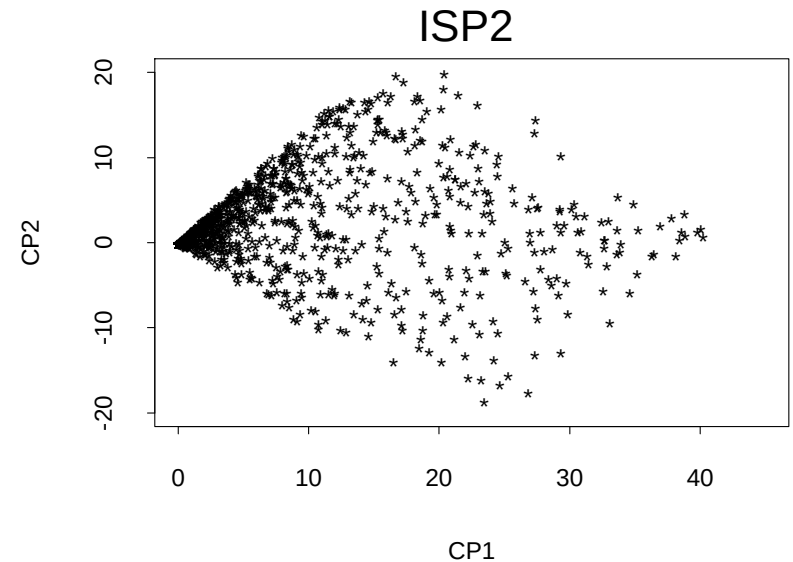
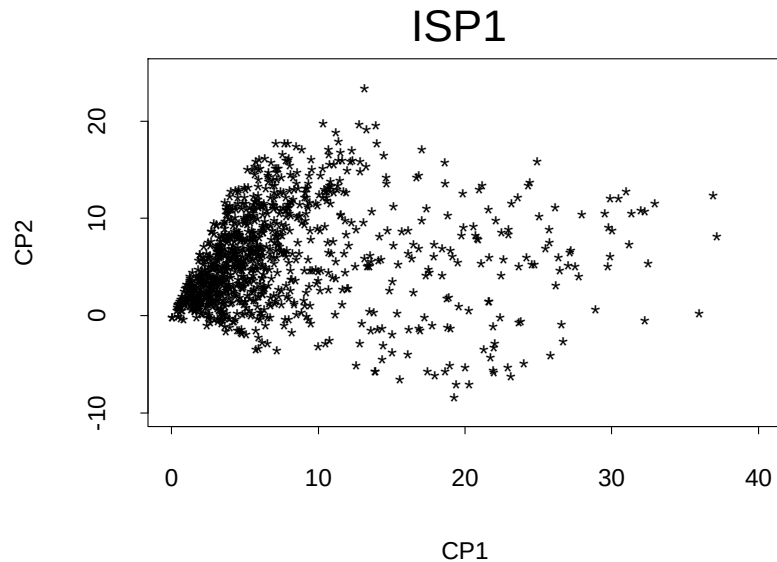
CP_2 – Contraste entre o taxas de transf. de manhã e de tar

2. Análise em componentes principais (cont.)

Cada utilizador pode ser representado nestes novos eixos de coordenadas:

2. Análise em componentes principais (cont.)

Cada utilizador pode ser representado nestes novos eixos de coordenadas:



3. *Análise de clusters*

Objectivo:

3. Análise de *clusters*

Objectivo: Encontrar grupos de indivíduos semelhantes, *i.e.*, grupos de indivíduos com um padrão de utilização da Internet ao longo do dia (Sábado) semelhante.

3. *Análise de clusters*

Objectivo: Encontrar grupos de indivíduos semelhantes, *i.e.*, grupos de indivíduos com um padrão de utilização da Internet ao longo do dia (Sábado) semelhante.

Consideraram-se dois métodos para formar clusters:

3. Análise de *clusters*

Objectivo: Encontrar grupos de indivíduos semelhantes, *i.e.*, grupos de indivíduos com um padrão de utilização da Internet ao longo do dia (Sábado) semelhante.

Consideraram-se dois métodos para formar clusters:

1. Aglomerativo: À partida, cada indivíduo é um *cluster*. O método vai agrupando *clusters* até obter um único grupo contendo todos os indivíduos.

3. Análise de *clusters*

Objectivo: Encontrar grupos de indivíduos semelhantes, *i.e.*, grupos de indivíduos com um padrão de utilização da Internet ao longo do dia (Sábado) semelhante.

Consideraram-se dois métodos para formar clusters:

1. Aglomerativo: À partida, cada indivíduo é um *cluster*. O método vai agrupando *clusters* até obter um único grupo contendo todos os indivíduos.

As medições tem natureza contínua logo a medida de proximidade escolhida é a **distância euclidiana**.

3. Análise de *clusters*

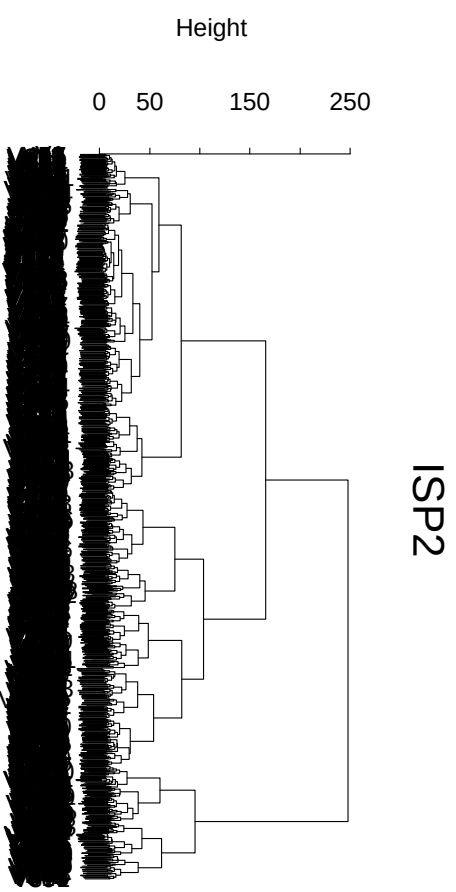
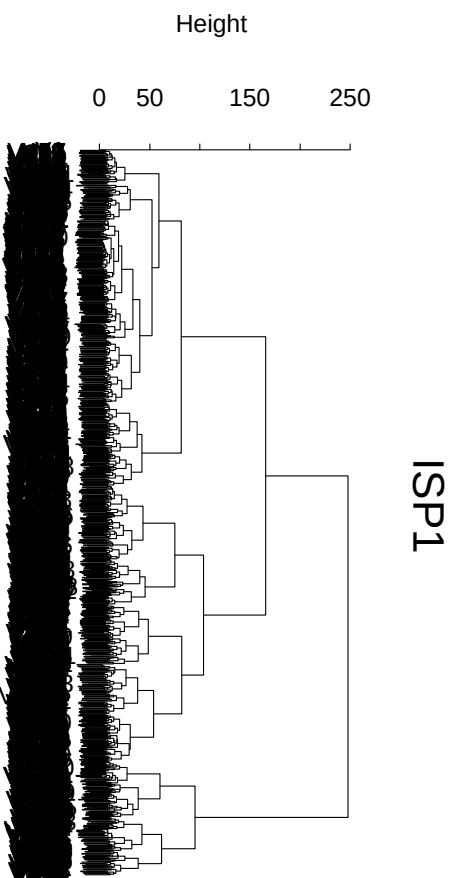
Objectivo: Encontrar grupos de indivíduos semelhantes, *i.e.*, grupos de indivíduos com um padrão de utilização da Internet ao longo do dia (Sábado) semelhante.

Consideraram-se dois métodos para formar clusters:

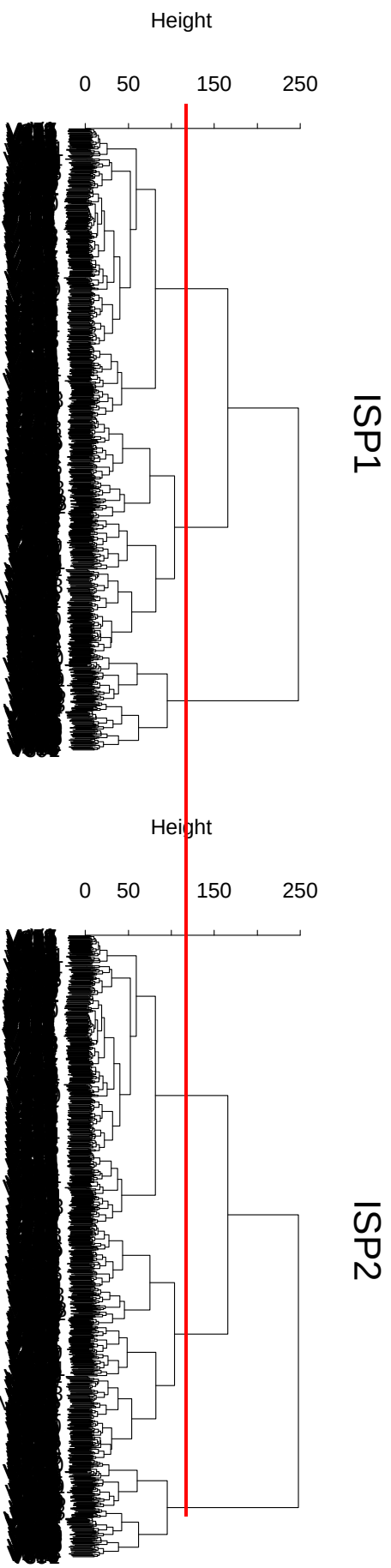
1. Aglomerativo: À partida, cada indivíduo é um *cluster*. O método vai agrupando *clusters* até obter um único grupo contendo todos os indivíduos.

As medições tem natureza contínua logo a medida de proximidade escolhida é a **distância euclidiana**. O critério de formação de *clusters* foi **método Ward** (*incremental sums of squares measure*).

3. Análise de *clusters* - Dendogramas (cont.)



3. Análise de *clusters* - Dendogramas (cont.)



-
-
-

3. *Análise de clusters - Medoids*

2. Medoids:

3. Análise de *clusters* - *Medoids*

2. Medoids: Método de partição, onde o número de *clusters* é escolhido à partida ($k = 3$).

3. Análise de *clusters* - *Medoids*

2. Medoids: Método de partição, onde o número de *clusters* é escolhido à partida ($k = 3$).

- Escolha dos *medoids*: k objectos escolhidos de forma a que distância de cada objecto ao *medoid* mais próximo é mínima.

3. Análise de *clusters* - *Medoids*

2. Medoids: Método de partição, onde o número de *clusters* é escolhido à partida ($k = 3$).

- Escolha dos *medoids*: k objectos escolhidos de forma a que distância de cada objecto ao *medoid* mais próximo é mínima.
- Cada objecto é atribuído ao *cluster* v_i se o *medoid* m_{v_i} está mais próximo do objecto i do que qualquer outro *medoid*.

3. Análise de *clusters* - *Medoids*

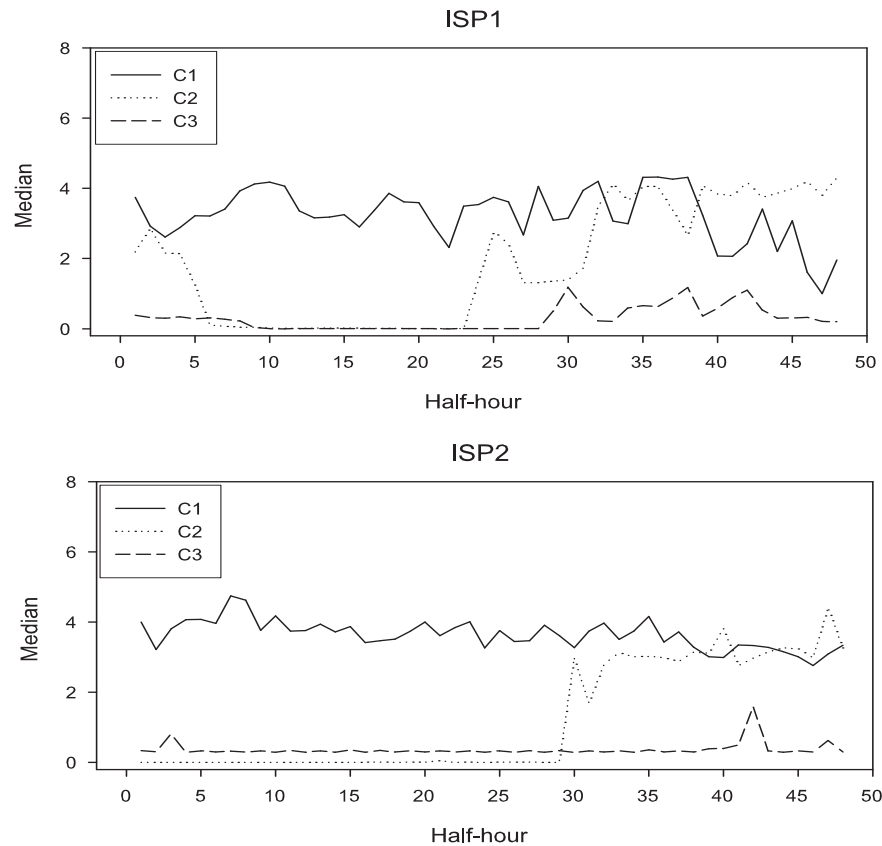
2. Medoids: Método de partição, onde o número de *clusters* é escolhido à partida ($k = 3$).

- Escolha dos *medoids*: k objectos escolhidos de forma a que distância de cada objecto ao *medoid* mais próximo é mínima.
- Cada objecto é atribuído ao *cluster* v_i se o *medoid* m_{v_i} está mais próximo do objecto i do que qualquer outro *medoid*.

Distância: **Euclidiana**

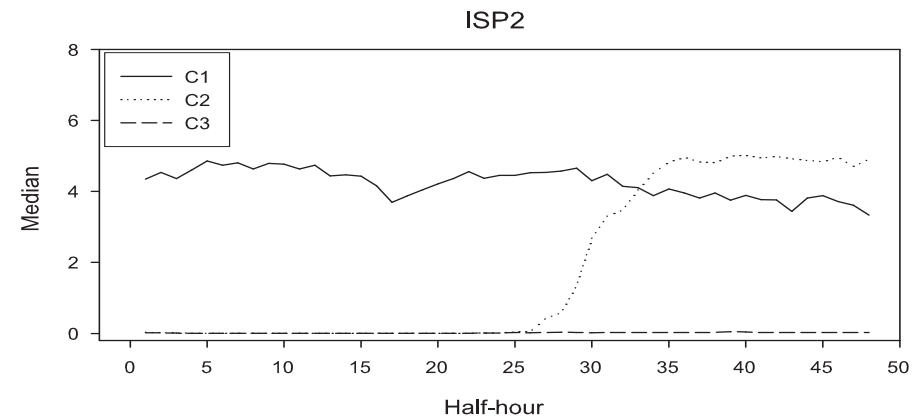
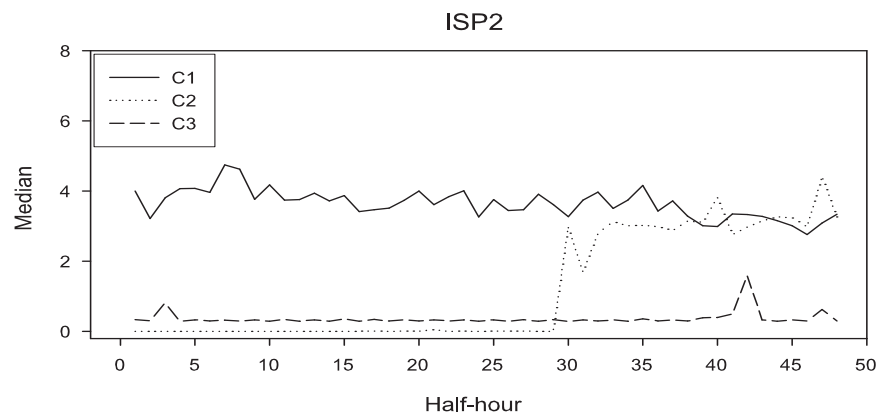
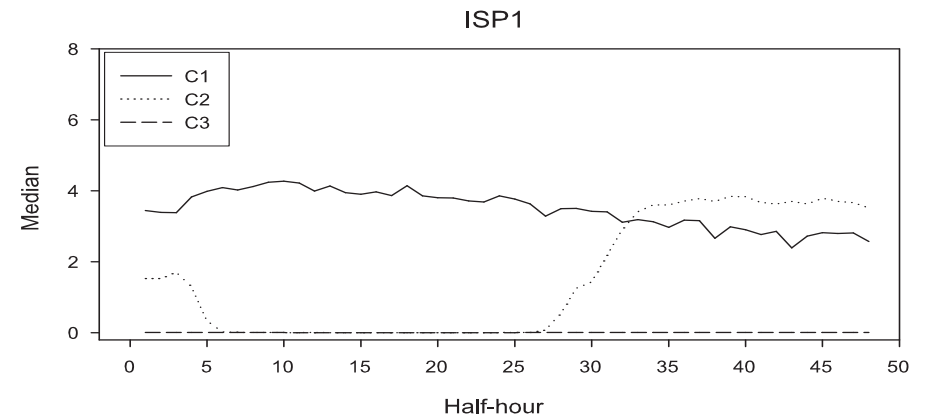
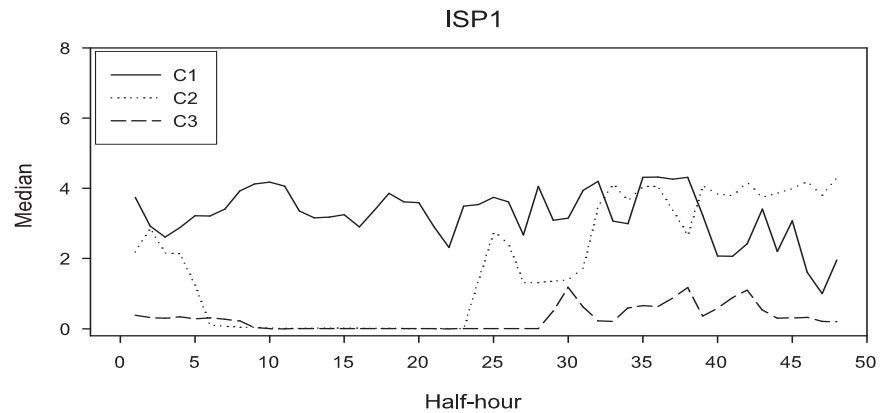
3. Análise de *clusters* - *Medoids*

Medoids

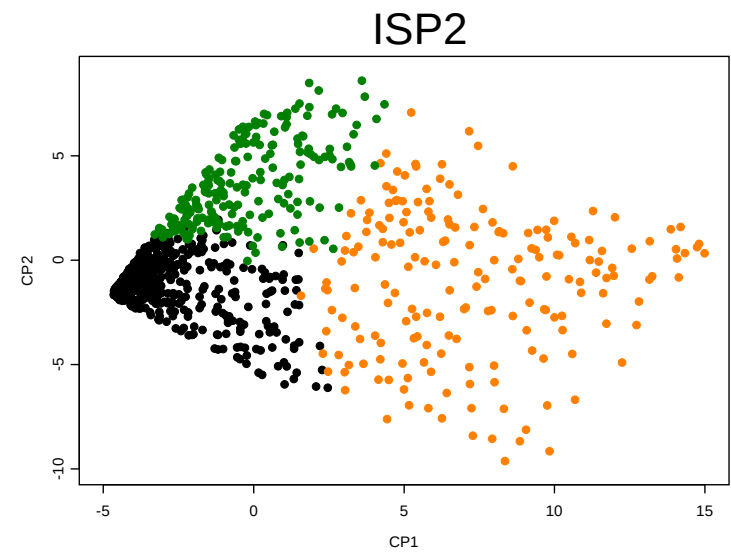
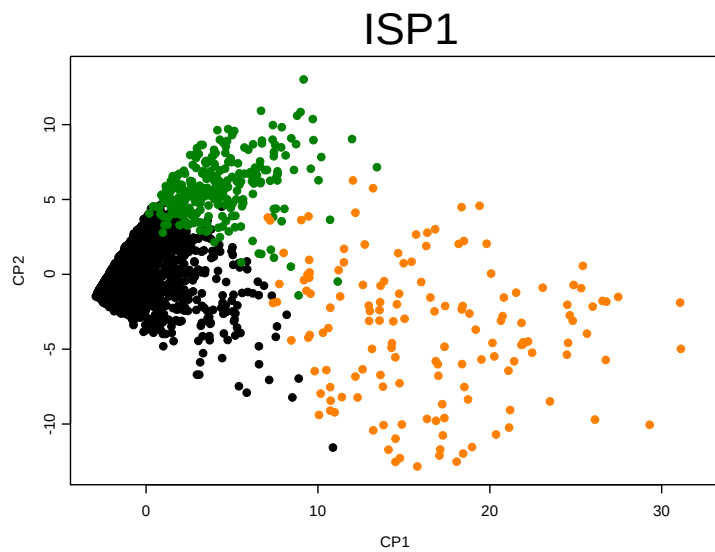


3. Análise de *clusters* - *Medoids*

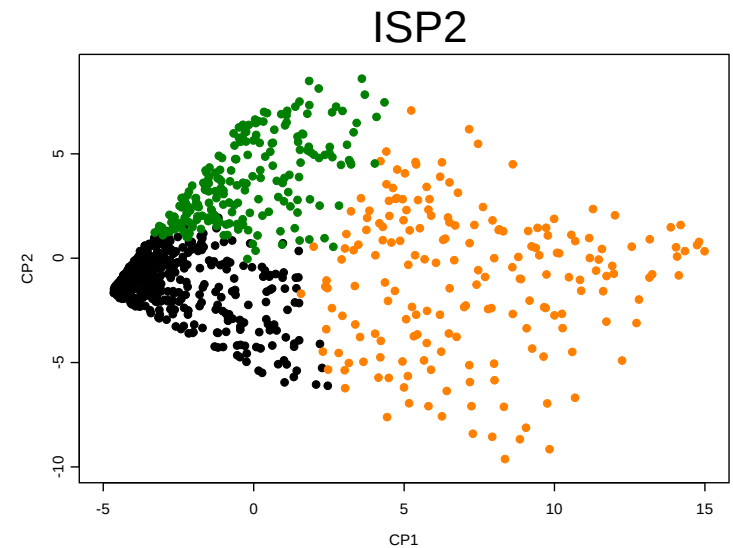
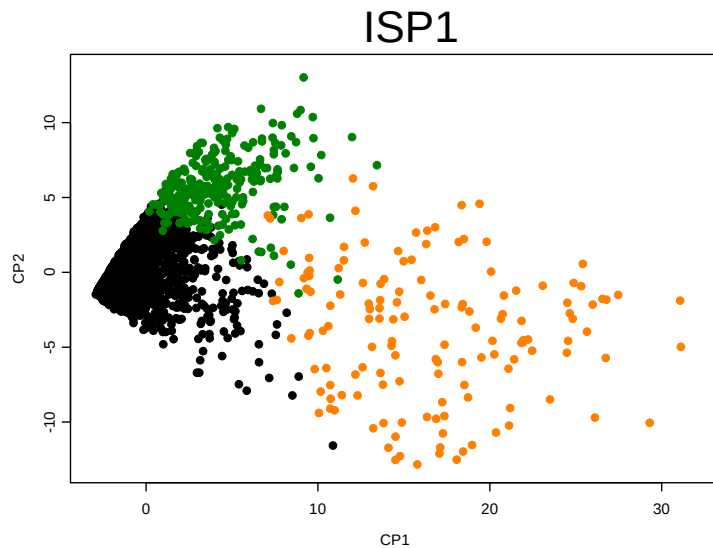
Medoids



3. Interpretation of clusters vs PC



3. Interpretation of clusters vs PC



<i>Cluster</i>	<i>Interpretação</i>
----------------	----------------------

- | | |
|----|--|
| C1 | Taxas de transf. elevadas ao longo de todo o dia |
| C2 | Taxas de transf. baixas/elevadas de manhã/tarde |
| C3 | Taxas de transf. baixas ao longo de todo o dia |
-

3. Medoids - Interpretação

Cluster	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
C1	4.23%	4.87%	18.52%	20.00%
C2	7.75%	22.29%	14.17%	21.94%
C3	88.02%	72.84%	67.32%	58.06%

3. Concordância entre as duas análise de *clusters*

Med.	Aglom. ISP1			
	C1	C2	C3	
C1	136	8	1	145
C2	6	258	2	266
C3	25	499	2497	3021
	167	765	2500	3432

Med.	Aglom. ISP2			
	C1	C2	C3	
C1	137	23	2	175
C2	8	104	12	192
C3	30	65	494	508
	162	124	589	875

3. Concordância entre as duas análise de *clusters*

Med.	Aglom. ISP1			
	C1	C2	C3	
C1	136	8	1	145
C2	6	258	2	266
C3	25	499	2497	3021
	167	765	2500	3432

Med.	Aglom. ISP2			
	C1	C2	C3	
C1	137	23	2	175
C2	8	104	12	192
C3	30	65	494	508
	162	124	589	875

Percentagem de casos concordantes entre atribuição a **3 clusters** feita por *Medoids* e Aglomerativo:

ISP1 : 84.2% *ISP2* : 84.0%

4. Análise discriminante

Objectivo:

4. Análise discriminante

Objectivo:

Encontrar uma regra para classificar um indivíduo, num de vários grupos pré-definidos, com base num conjunto de variáveis.

4. Análise discriminante

Objectivo:

Encontrar uma regra para classificar um indivíduo, num de vários grupos pré-definidos, com base num conjunto de variáveis.

Grupo: Perfil a que cada indivíduo é atribuído pela análise de *clusters* (*Medoids*).

4. Análise discriminante

Objectivo:

Encontrar uma regra para classificar um indivíduo, num de vários grupos pré-definidos, com base num conjunto de variáveis.

Grupo: Perfil a que cada indivíduo é atribuído pela análise de *clusters* (*Medoids*).

Com base no seus consumos ao longo do dia procura-se re-classificar o indivíduo num dos três perfis considerados.

4. Análise discriminante

Objectivo:

Encontrar uma regra para classificar um indivíduo, num de vários grupos pré-definidos, com base num conjunto de variáveis.

Grupo: Perfil a que cada indivíduo é atribuído pela análise de *clusters* (*Medoids*).

Com base no seus consumos ao longo do dia procura-se re-classificar o indivíduo num dos três perfis considerados.

Se a regra discriminante apresentar taxas de erro baixas estaremos a validar a divisão em *clusters* feita inicialmente.

4. Análise discriminante (cont.)

Taxa erro	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
Aparente	1.31%	9.47%	3.09%	6.74%
Leave-one-out	1.81%	9.99%	5.71%	9.49%
Valid-Cruzada	1.78%	10.05%	6.59%	9.57%
Dupla Valid-Cruz.	2.18%	10.02%	7.77%	10.86%

4. Análise discriminante (cont.)

Taxa erro	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
Aparente	1.31%	9.47%	3.09%	6.74%
Leave-one-out	1.81%	9.99%	5.71%	9.49%
Valid-Cruzada	1.78%	10.05%	6.59%	9.57%
Dupla Valid-Cruz.	2.18%	10.02%	7.77%	10.86%

Em geral, as taxas de erro são baixas, o que valida a estrutura de *clusters* proposta.

4. Análise discriminante (cont.)

Taxa erro	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
Aparente	1.31%	9.47%	3.09%	6.74%
Leave-one-out	1.81%	9.99%	5.71%	9.49%
Valid-Cruzada	1.78%	10.05%	6.59%	9.57%
Dupla Valid-Cruz.	2.18%	10.02%	7.77%	10.86%

Em geral, as taxas de erro são baixas, o que valida a estrutura de *clusters* proposta. Note que as partições obtidas pelos Medoids apresentam menores taxas de erro.

4. Análise discriminante (cont.)

Taxa erro	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
Aparente	1.31%	9.47%	3.09%	6.74%
Leave-one-out	1.81%	9.99%	5.71%	9.49%
Valid-Cruzada	1.78%	10.05%	6.59%	9.57%
Dupla Valid-Cruz.	2.18%	10.02%	7.77%	10.86%

Em geral, as taxas de erro são baixas, o que valida a estrutura de *clusters* proposta. Note que as partições obtidas pelos Medoids apresentam menores taxas de erro. Por esta razão serão as partições consideradas.

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

- Número de bytes recebidos por utilizador (**Tamanho**);

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

- Número de bytes recebidos por utilizador (**Tamanho**);
- **Duração** por utilizador;

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

- Número de bytes recebidos por utilizador (**Tamanho**);
- **Duração** por utilizador;
- **Taxa** de transferência por utilizador

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

- Número de bytes recebidos por utilizador (**Tamanho**);
- **Duração** por utilizador;
- **Taxa** de transferência por utilizador

em cada um dos clusters: **C1**, **C2** e C3.

5. Caracterização dos *clusters*

Objectivo: Analisar cada *cluster* com base em características do tráfego que não foram utilizadas na análise de *clusters*

Em particular estamos interessados em perceber como se caracterizam

- Número de bytes recebidos por utilizador (**Tamanho**);
- **Duração** por utilizador;
- **Taxa** de transferência por utilizador

em cada um dos clusters: **C1**, **C2** e C3.

Analisaremos apenas a partição obtida pelo métodos dos **medoids**.

6. Caracterização dos *clusters*

Usando um teste não-paramétrico, concluiu-se que se deve

6. Caracterização dos *clusters*

Usando um teste não-paramétrico, concluiu-se que se deve

Rejeitar $H_0 : \mu_1 = \mu_2 = \mu_3$

6. Caracterização dos *clusters*

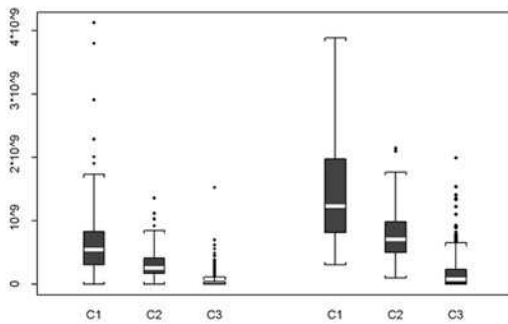
Usando um teste não-paramétrico, concluiu-se que se deve

Rejeitar $H_0 : \mu_1 = \mu_2 = \mu_3$

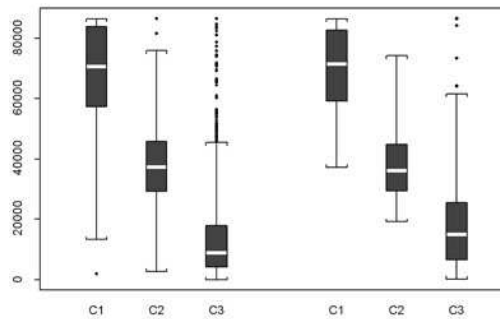
para Tamanho, Taxas de Transferência e Duração

6. Caracterização dos *clusters*

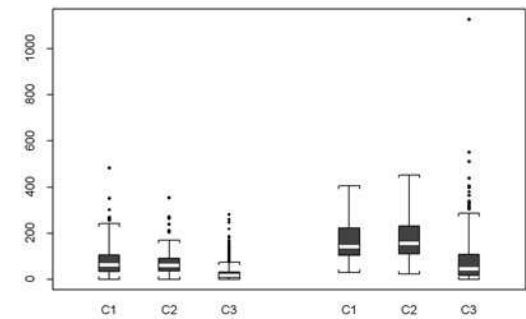
Tamanho



Duração

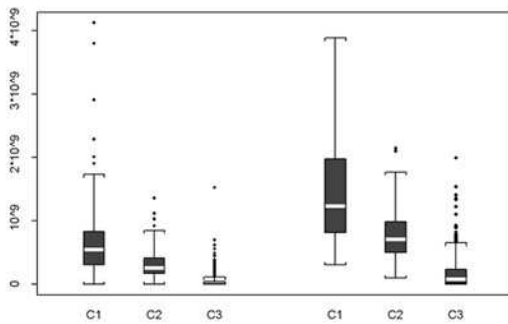


Taxa

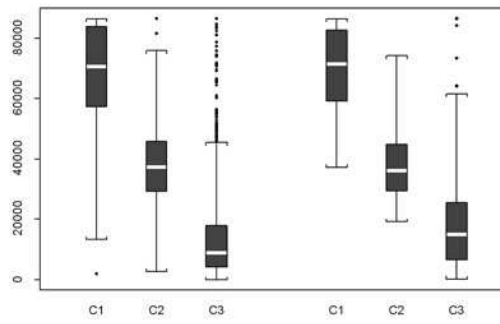


6. Caracterização dos *clusters*

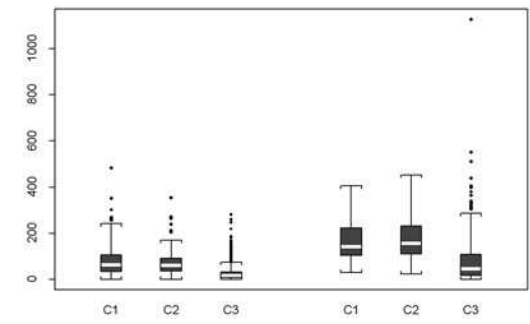
Tamanho



Duração



Taxa

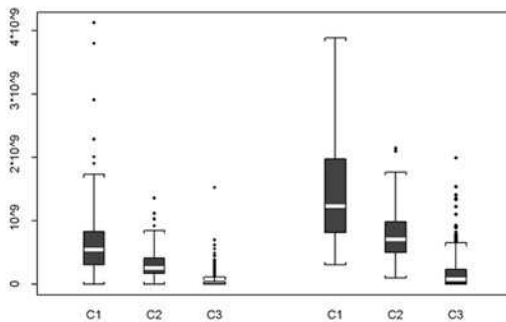


ISP2: $\mu_1 > \mu_2 > \mu_3$

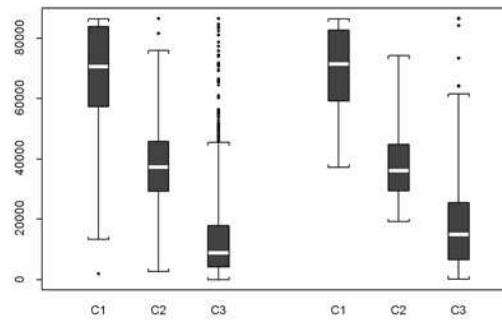
ISP1: $\mu_1 = \mu_2 > \mu_3$

6. Caracterização dos *clusters*

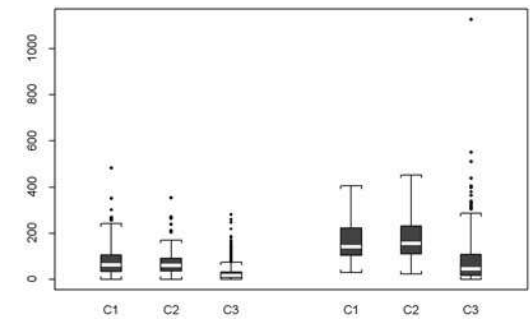
Tamanho



Duração



Taxa



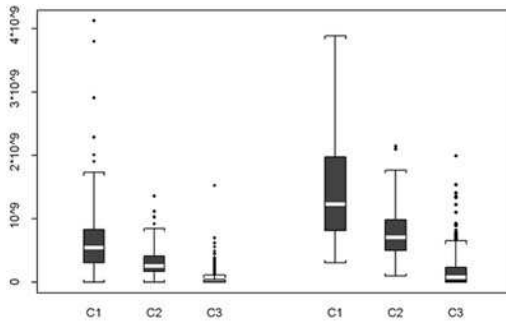
ISP2: $\mu_1 > \mu_2 > \mu_3$

ISP1: $\mu_1 = \mu_2 > \mu_3$

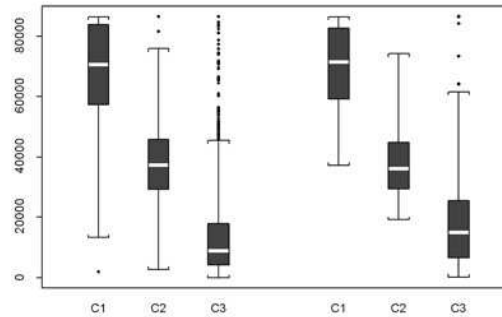
$\mu_1 > \mu_2 > \mu_3$

6. Caracterização dos *clusters*

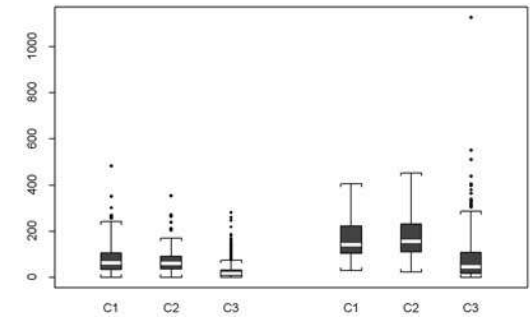
Tamanho



Duração



Taxa



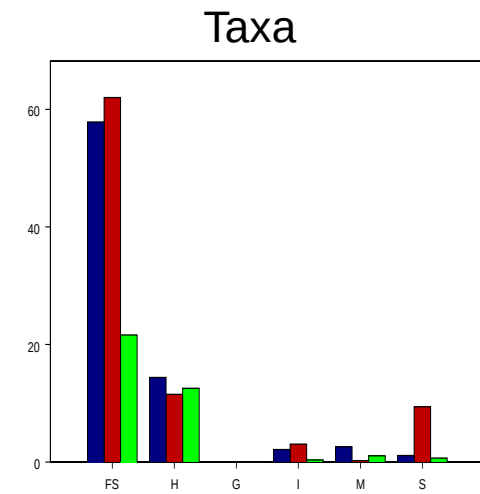
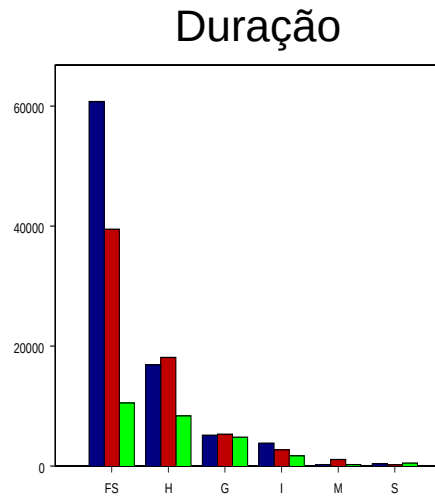
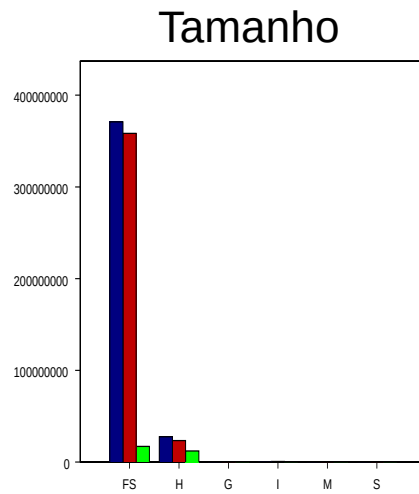
ISP2: $\mu_1 > \mu_2 > \mu_3$

$\mu_1 > \mu_2 > \mu_3$

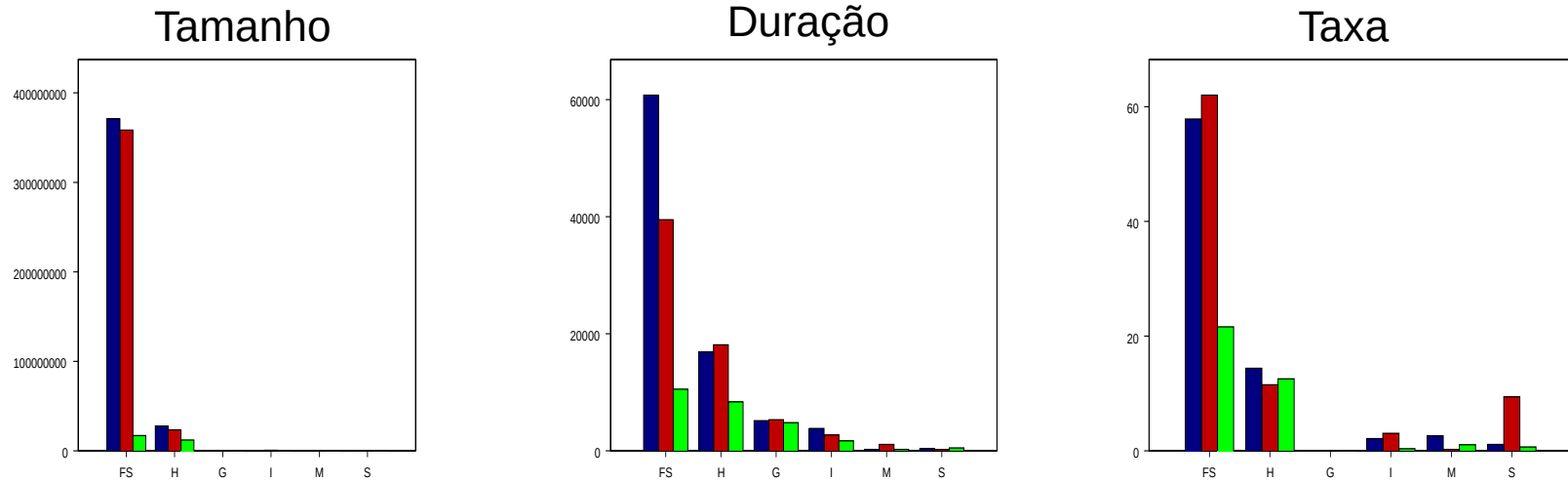
$\mu_1 = \mu_2 > \mu_3$

ISP1: $\mu_1 = \mu_2 > \mu_3$

6. Caracterização dos *clusters* de ISP2 - Aplicações

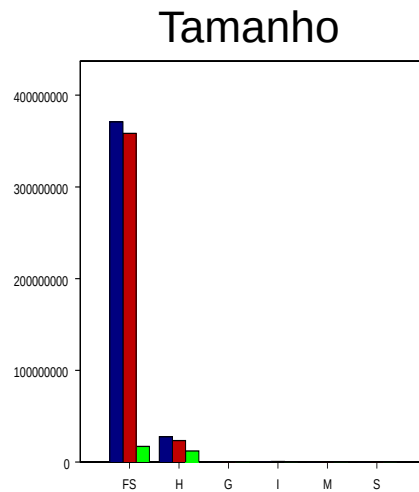


6. Caracterização dos *clusters* de ISP2 - Aplicações

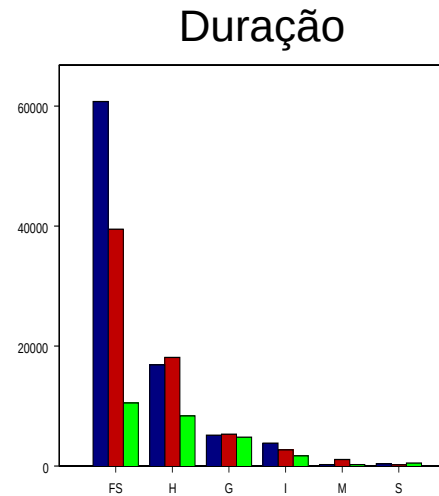


FS, H: $\mu_1 = \mu_2 > \mu_3$

6. Caracterização dos *clusters* de ISP2 - Aplicações

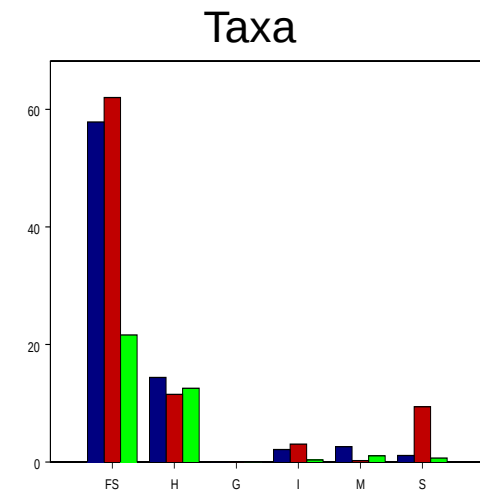


FS, H: $\mu_1 = \mu_2 > \mu_3$

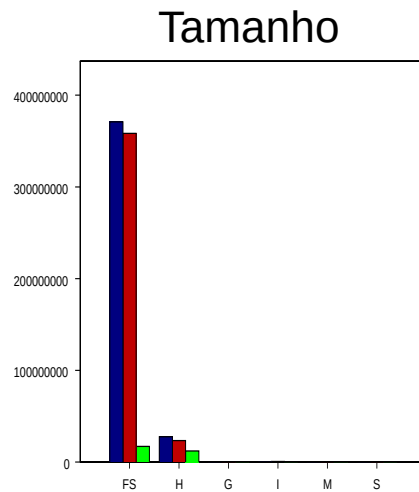


FS: $\mu_1 > \mu_2 > \mu_3$

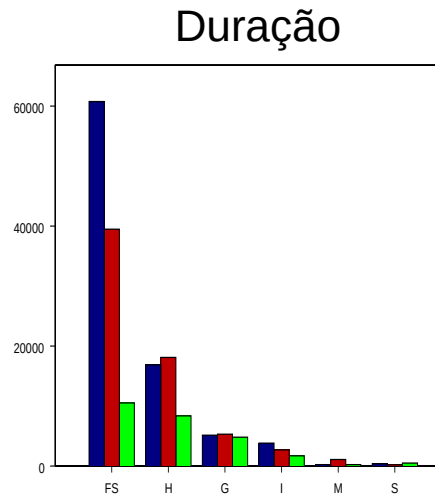
H: $\mu_1 = \mu_2 > \mu_3$



6. Caracterização dos *clusters* de ISP2 - Aplicações

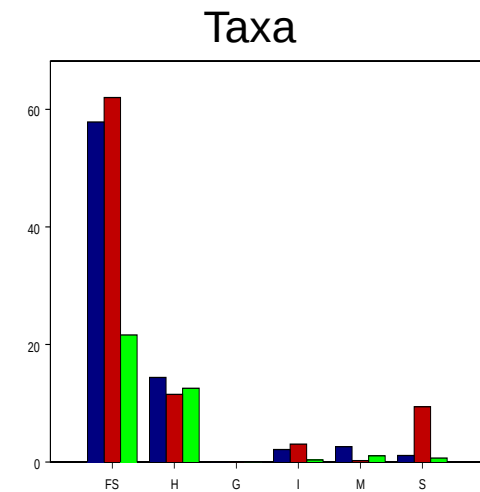


FS, H: $\mu_1 = \mu_2 > \mu_3$



FS: $\mu_1 > \mu_2 > \mu_3$

H: $\mu_1 = \mu_2 > \mu_3$



FS: $\mu_1 = \mu_2 > \mu_3$

Conclusões I

- **Análise de Cluster:** Obtivemos uma partição em 3 grupos com perfis de utilização da Internet semelhantes.

Conclusões I

- **Análise de Cluster:** Obtivemos uma partição em 3 grupos com perfis de utilização da Internet semelhantes.
- Um utilizador típico de cada cluster é caracterizado por:

Conclusões I

- **Análise de Cluster:** Obtivemos uma partição em 3 grupos com perfis de utilização da Internet semelhantes.
- Um utilizador típico de cada cluster é caracterizado por:
 - **C1:** Taxas de transf. elevadas ao longo de todo o dia.

Conclusões I

- **Análise de Cluster:** Obtivemos uma partição em 3 grupos com perfis de utilização da Internet semelhantes.
- Um utilizador típico de cada cluster é caracterizado por:
 - **C1:** Taxas de transf. elevadas ao longo de todo o dia.
 - **C2:** Taxas de transf. baixas/elevadas de manhã/tarde.

Conclusões I

- **Análise de Cluster:** Obtivemos uma partição em 3 grupos com perfis de utilização da Internet semelhantes.
- Um utilizador típico de cada cluster é caracterizado por:
 - **C1:** Taxas de transf. elevadas ao longo de todo o dia.
 - **C2:** Taxas de transf. baixas/elevadas de manhã/tarde.
 - **C3:** Taxas de transf. baixas ao longo de todo o dia.

Conclusões I

- Utilizadores do perfil **C1** e **C2** usam **File Sharing** a taxas mais elevadas e durante períodos mais longos que C3.

Conclusões I

- Utilizadores do perfil **C1** e **C2** usam **File Sharing** a taxas mais elevadas e durante períodos mais longos que C3.
- Os utilizadores de **C1** tendem a usar a Internet de forma **mais intensiva** do que os de **C2**.

Conclusões I

- Utilizadores do perfil **C1** e **C2** usam **File Sharing** a taxas mais elevadas e durante períodos mais longos que C3.
- Os utilizadores de **C1** tendem a usar a Internet de forma **mais intensiva** do que os de **C2**.
- O tráfego associado aos utilizadores agrupados em C3 caracteriza-se por menores durações e menos bytes recebidos de **HTTP** que **C1** ou **C2**.

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Baseado em: - Medições de taxas de transferência, medidas durante uma semana. ($n = 5588$, $p = 346$ períodos de 30min, de Terça a Terça)

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Baseado em: - Medições de taxas de transferência, medidas durante uma semana. ($n = 5588$, $p = 346$ períodos de 30min, de Terça a Terça)

Útil para:

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Baseado em: - Medições de taxas de transferência, medidas durante uma semana. ($n = 5588$, $p = 346$ períodos de 30min, de Terça a Terça)

Útil para:

- Entender a natureza do tráfego da Internet.

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Baseado em: - Medições de taxas de transferência, medidas durante uma semana. ($n = 5588$, $p = 346$ períodos de 30min, de Terça a Terça)

Útil para:

- Entender a natureza do tráfego da Internet.
- Detectar perfis semanais de utilização da Internet.

Análise da estrutura do tráfego da Internet

Objectivos - Decomposição do tráfego associado a cada utilizador em componentes com características marcadamente diferentes.

Baseado em: - Medições de taxas de transferência, medidas durante uma semana. ($n = 5588$, $p = 346$ períodos de 30min, de Terça a Terça)

Útil para:

- Entender a natureza do tráfego da Internet.
- Detectar perfis semanais de utilização da Internet.

Transformações: Logaritmizar e centar.

Análise da estrutura do tráfego da Internet

Seja X a matriz dos dados $n \times p$ ($n = 5588$ e $p = 346$), que se decompõe usando a decomposição em valores e vectores singulares:

Análise da estrutura do tráfego da Internet

Seja X a matriz dos dados $n \times p$ ($n = 5588$ e $p = 346$), que se decompõe usando a decomposição em valores e vectores singulares:

$$X = U \Lambda V^t$$

Análise da estrutura do tráfego da Internet

Seja X a matriz dos dados $n \times p$ ($n = 5588$ e $p = 346$), que se decompõe usando a decomposição em valores e vectores singulares:

$$X = U \Lambda V^t$$

onde:

– $U = (u_1, \dots, u_p)$ é uma matriz de dimensão $n \times p$, tendo em colunas os primeiros p vectores próprios de XX^t

Análise da estrutura do tráfego da Internet

Seja X a matriz dos dados $n \times p$ ($n = 5588$ e $p = 346$), que se decompõe usando a decomposição em valores e vectores singulares:

$$X = U \Lambda V^t$$

onde:

- $U = (u_1, \dots, u_p)$ é uma matriz de dimensão $n \times p$, tendo em colunas os primeiros p vectores próprios de XX^t
- $V = (v_1, \dots, v_p)$ é uma matriz de dimensão $p \times p$, tendo em colunas os p vectores próprios de $X^t X$

Análise da estrutura do tráfego da Internet

Seja $\mathbf{X}$ a matriz dos dados $n \times p$ ($n = 5588$ e $p = 346$), que se decompõe usando a decomposição em valores e vectores singulares:

$$\mathbf{X} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^t$$

onde:

- $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$ é uma matriz de dimensão $n \times p$, tendo em colunas os primeiros p vectores próprios de $\mathbf{X} \mathbf{X}^t$
- $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_p)$ é uma matriz de dimensão $p \times p$, tendo em colunas os p vectores próprios de $\mathbf{X}^t \mathbf{X}$
- $\mathbf{\Lambda} = \text{diag} \{ \lambda_1, \dots, \lambda_p \}$, onde $\lambda_1 \geq \dots \geq \lambda_p > 0$ são as raízes quadradas dos primeiros p valores próprios de $\mathbf{X} \mathbf{X}^t$.

Análise da estrutura do tráfego da Internet

Uma vez que X é centrada V contem os **loadings das componentes principais** associadas a S (matriz de covariâncias amostral).

Análise da estrutura do tráfego da Internet

Uma vez que X é centrada V contem os **loadings das componentes principais** associadas a S (matriz de covariâncias amostral).

O tráfego associado ao i -ésimo utilizador (i -ésima linha de X) pode escrever-se como uma combinação linear das colunas de V , que se designa por **fluxos próprios** (*eigenflows*).

Análise da estrutura do tráfego da Internet

Uma vez que X é centrada V contem os **loadings das componentes principais** associadas a S (matriz de covariâncias amostral).

O tráfego associado ao i -ésimo utilizador (i -ésima linha de X) pode escrever-se como uma combinação linear das colunas de V , que se designa por **fluxos próprios** (*eigenflows*).

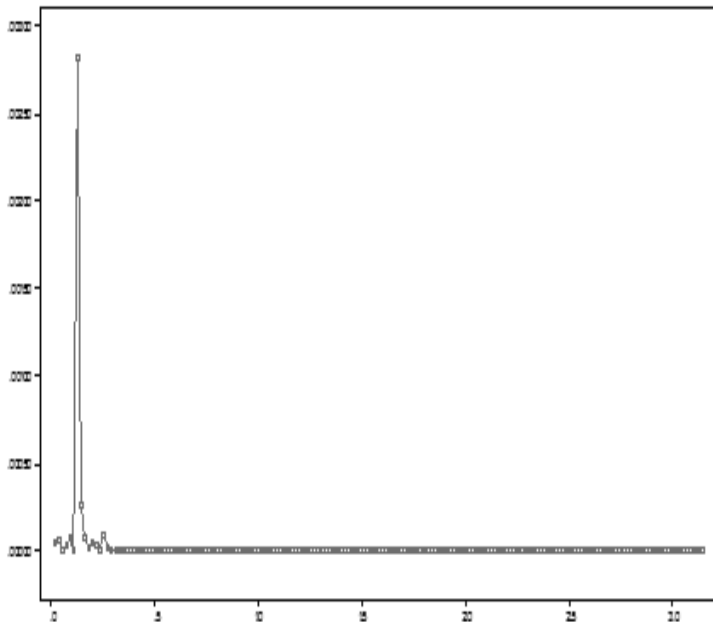
$$x_i = V\delta_i = \delta_{i1}v_1 + \dots + \delta_{ip}v_p$$

Classificação dos fluxos próprios

Cada fluxo próprio é classificado numa de três categorias: periódicos (**d-eigenflows**), ruído (**n-eigenflows**) e picos (**s-eigenflows**).

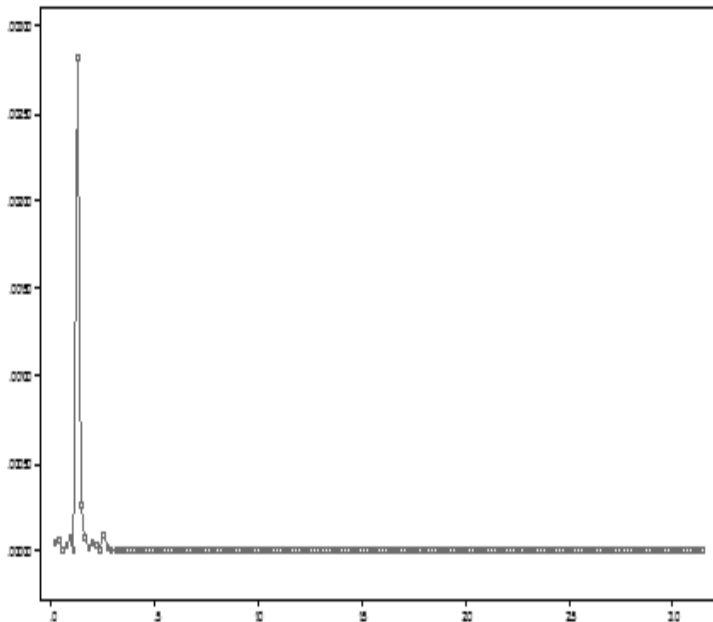
Classificação dos fluxos próprios

Cada fluxo próprio é classificado numa de três categorias: periódicos (**d-eigenflows**), ruído (**n-eigenflows**) e picos (**s-eigenflows**).



Classificação dos fluxos próprios

Cada fluxo próprio é classificado numa de três categorias: periódicos (**d-eigenflows**), ruído (**n-eigenflows**) e picos (**s-eigenflows**).



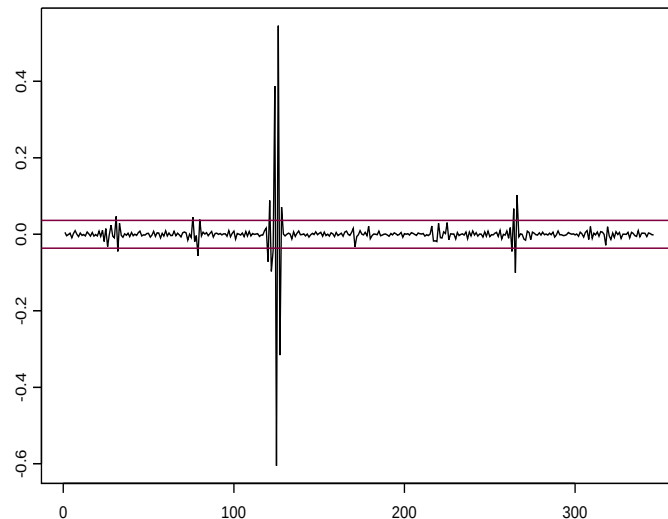
Os d-eigenflows são fluxos próprios cujo periodograma revela periodicidade, usualmente com picos correspondentes a períodos de 12 ou 24 horas.

Classificação dos fluxos próprios

Um fluxo próprio é classificado como um **s-eigenflow** se apresentar pelo menos um “grande” pico, i.e. se pelo menos um das suas entradas estiver fora do intervalo:

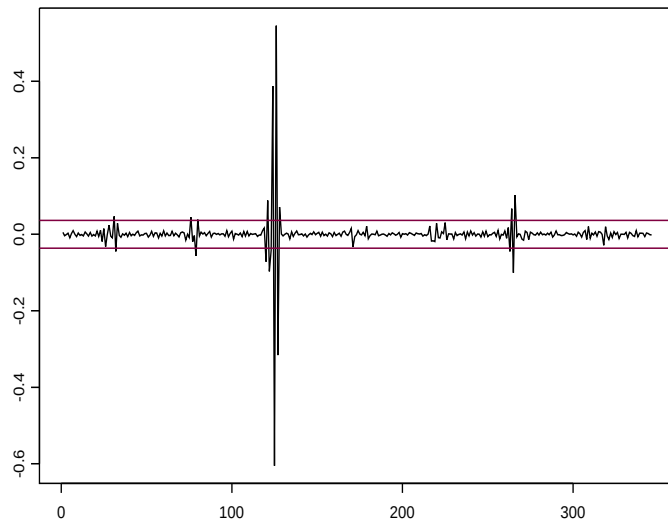
Classificação dos fluxos próprios

Um fluxo próprio é classificado como um **s-eigenflow** se apresentar pelo menos um “grande” pico, i.e. se pelo menos um das suas entradas estiver fora do intervalo:



Classificação dos fluxos próprios

Um fluxo próprio é classificado como um **s-eigenflow** se apresentar pelo menos um “grande” pico, i.e. se pelo menos um das suas entradas estiver fora do intervalo:



$$[med(\mathbf{v}_j) - 6MAD(\mathbf{v}_j), med(\mathbf{v}_j) + 6MAD(\mathbf{v}_j)],$$

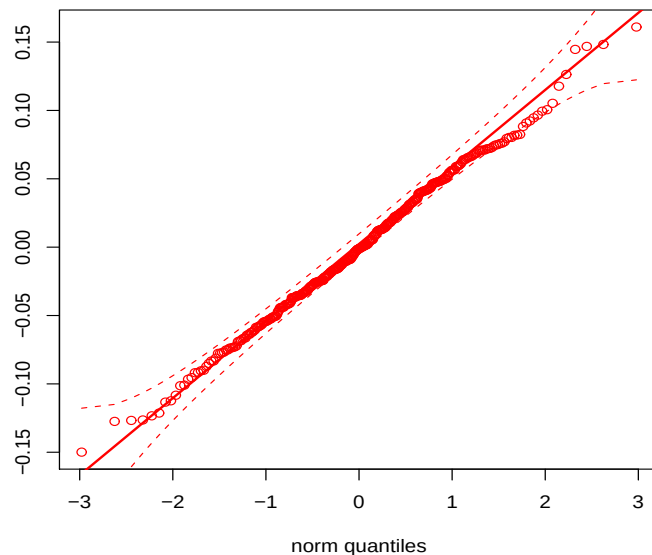
onde **MAD** é um estimador robusto de dispersão.

Classificação dos fluxos próprios

Os **n-eigenflow** são os restantes fluxos próprios, tais que:

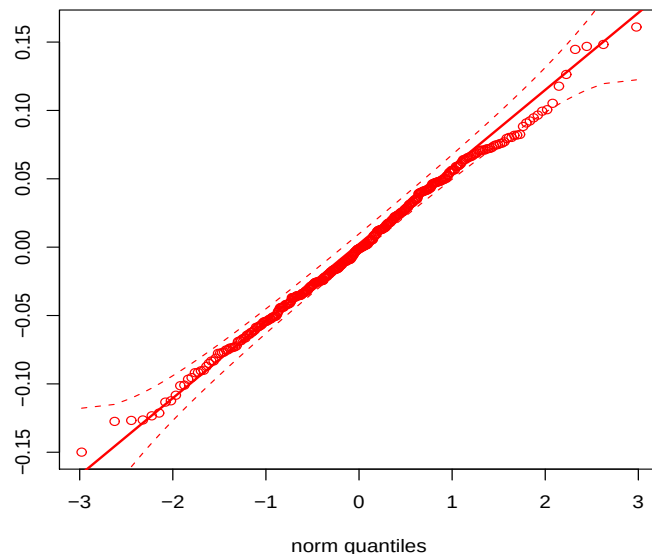
Classificação dos fluxos próprios

Os **n-eigenflow** são os restantes fluxos próprios, tais que:



Classificação dos fluxos próprios

Os **n-eigenflow** são os restantes fluxos próprios, tais que:



(i) Teste de chi-quadrado não rejeita a hipótese de terem distribuição normal,

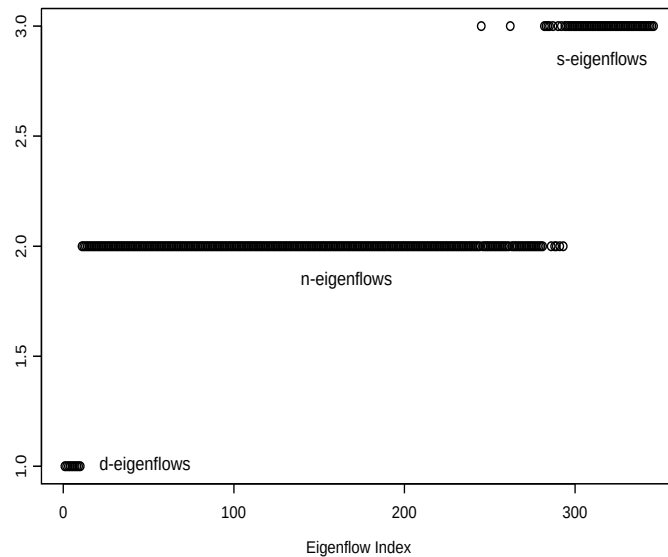
(ii) Rejeitando-se a hipótese de normalidade não foram classificados em nenhuma das categorias anteriores (34 em 346).

Classificação dos fluxos próprios

Da classificação dos fluxos próprios pode concluir-se que:

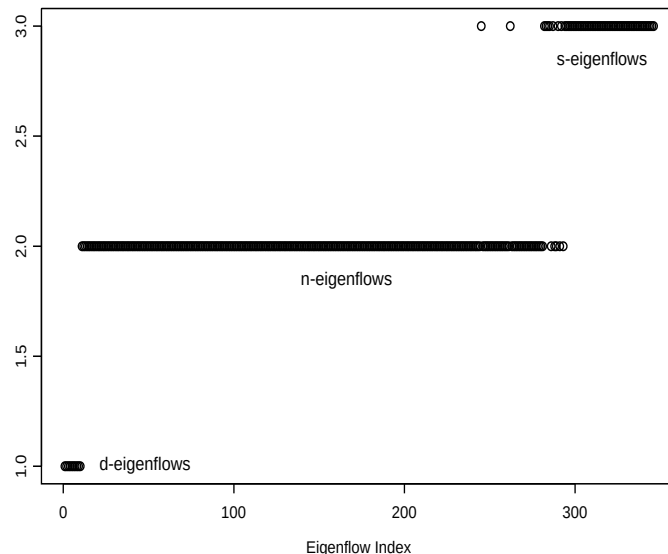
Classificação dos fluxos próprios

Da classificação dos fluxos próprios pode concluir-se que:



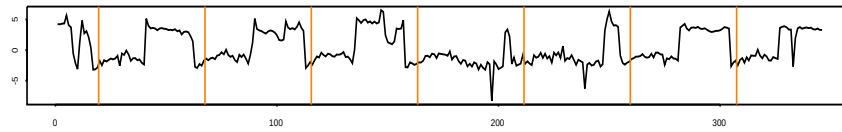
Classificação dos fluxos próprios

Da classificação dos fluxos próprios pode concluir-se que:

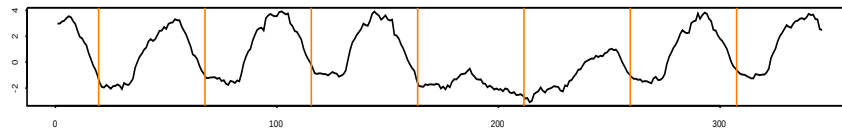


- (i) Primeiros 10 são **d-eigenflows** e explicam 52% da variabilidade total.
- (ii) Os últimos 62 são **s-eigenflows** e explicam 1% da variabilidade total.
- (iii) 247 fluxos próprios, classificados como **n-eigenflows**, explicam 47% da variabilidade total.

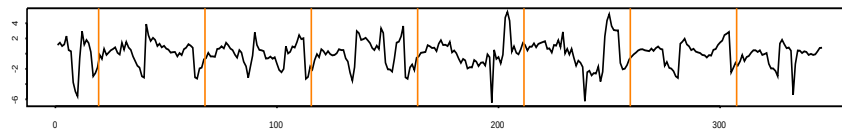
Classificação dos fluxos próprios



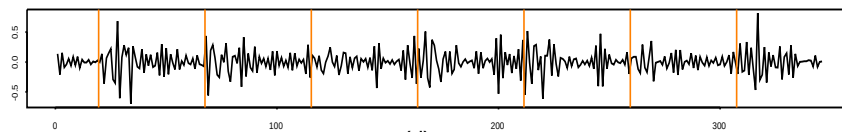
(a)



(b)

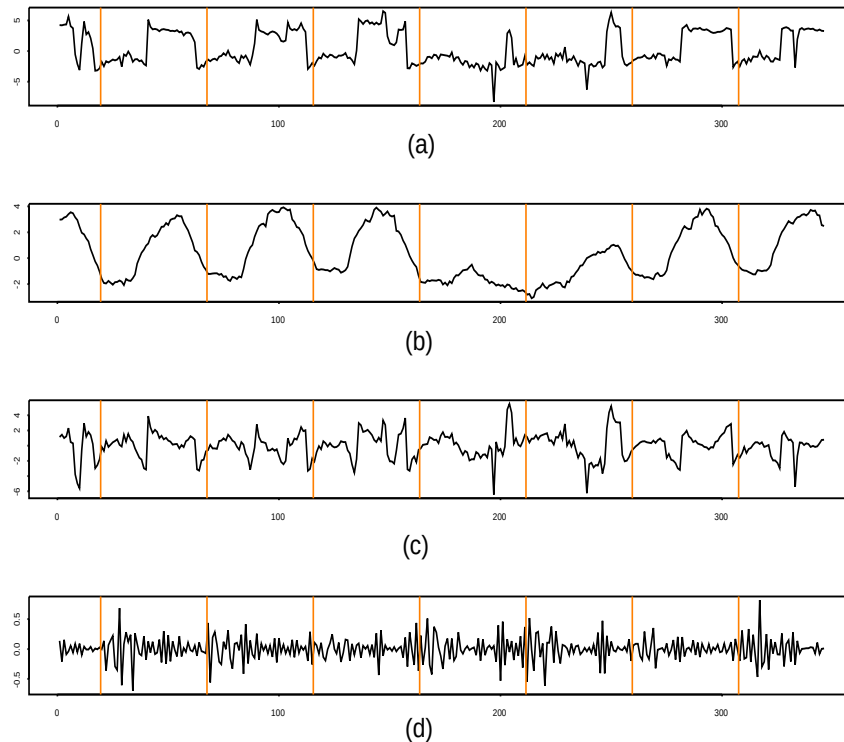


(c)



(d)

Classificação dos fluxos próprios



Uma vez que o tráfego associado a cada utilizador, x_i , pode ser escrito como uma combinação linear de fluxos próprios, x_i (a) pode ser decomposta em 3 componentes:

(b) periódica,
(c) ruído,
(d) picos.

Análise da estrutura do tráfego da Internet

A decomposição sugere que os padrões de utilização semanal da Internet comuns aos vários utilizadores são explicados, essencialmente, pela componente periódica.

Análise da estrutura do tráfego da Internet

A decomposição sugere que os padrões de utilização semanal da Internet comuns aos vários utilizadores são explicados, essencialmente, pela componente periódica.

Para analisar este pressuposto, truncaram-se os vectores resposta à componentes periódica, i.e.

Análise da estrutura do tráfego da Internet

A decomposição sugere que os padrões de utilização semanal da Internet comuns aos vários utilizadores são explicados, essencialmente, pela componente periódica.

Para analisar este pressuposto, truncaram-se os vectores resposta à componentes periódica, i.e.

$$\mathbf{x}_i^{tr} = \sum_{j=1}^{10} \delta_{ij} \mathbf{v}_j,$$

Análise da estrutura do tráfego da Internet

A decomposição sugere que os padrões de utilização semanal da Internet comuns aos vários utilizadores são explicados, essencialmente, pela componente periódica.

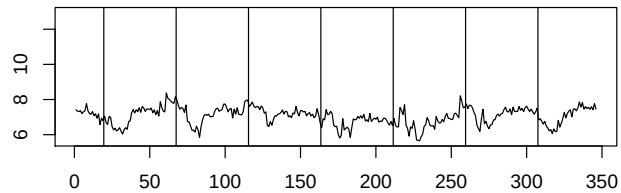
Para analisar este pressuposto, truncaram-se os vectores resposta à componentes periódica, i.e.

$$\mathbf{x}_i^{tr} = \sum_{j=1}^{10} \delta_{ij} \mathbf{v}_j,$$

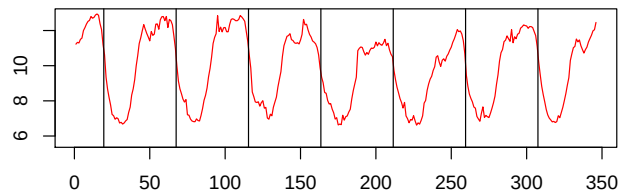
e fez-se análise de clusters baseada nos vectores resposta **completos** e **truncados**.

Análise de Clusters - $k=2$

Completo

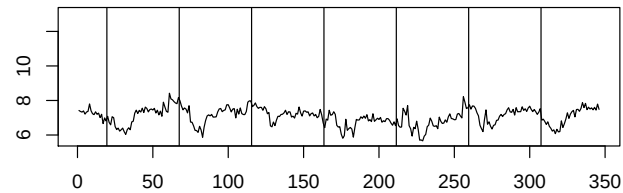


(a)

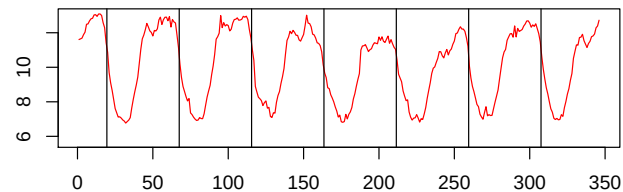


(b)

Truncado



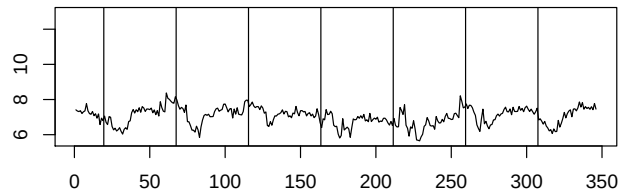
(a)



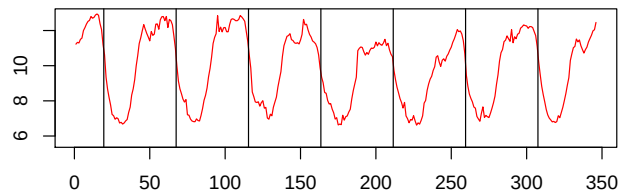
(b)

Análise de Clusters - $k=2$

Completo

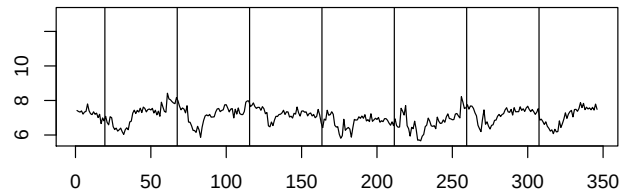


(a)

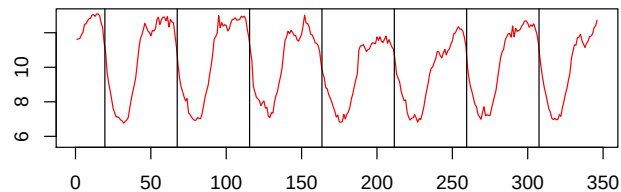


(b)

Truncado



(a)

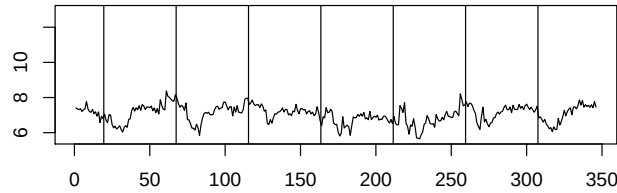


(b)

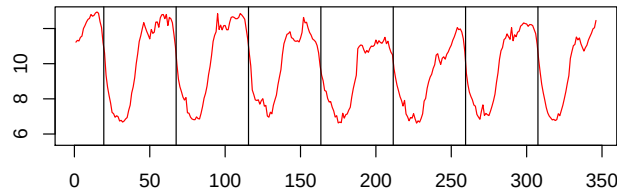
Taxa de concordância entre as duas partições: 96%

Análise de Clusters - k=2

Completo

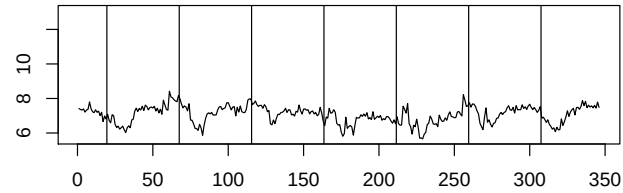


(a)

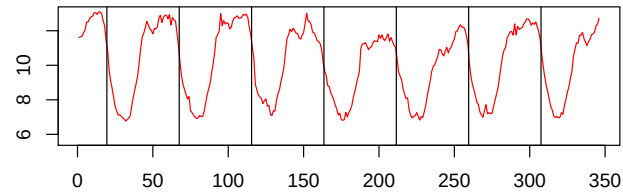


(b)

Truncado



(a)



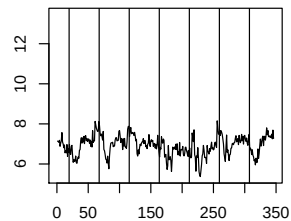
(b)

Taxa de concordância entre as duas partições: 96%

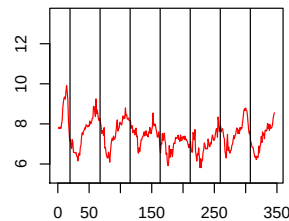
- **C1** - Taxas de transferência baixas ao longo de todo o dia.
- **C2** - Taxas de transferência baixas no início do dia e elevadas no final do dia e noite.

Análise de Clusters - $k=3$

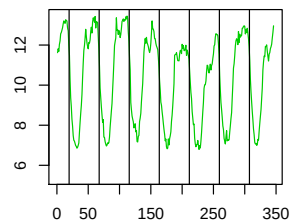
Completo



(a)

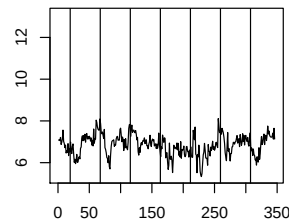


(b)

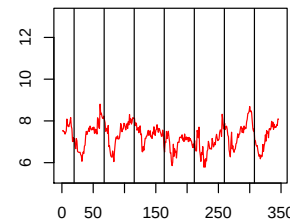


(c)

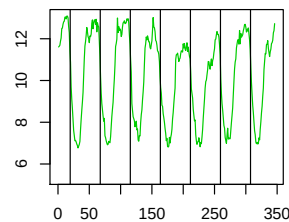
Truncado



(a)



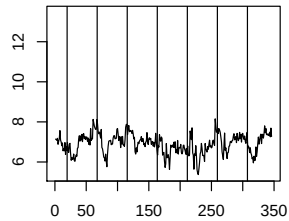
(b)



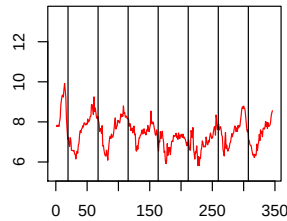
(c)

Análise de Clusters - k=3

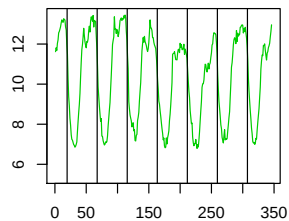
Completo



(a)

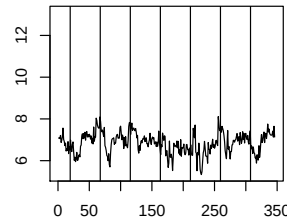


(b)

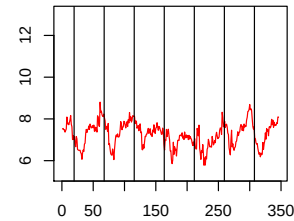


(c)

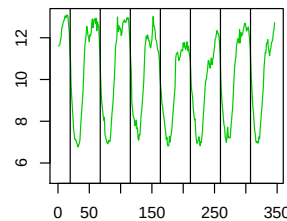
Truncado



(a)



(b)

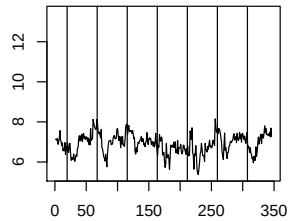


(c)

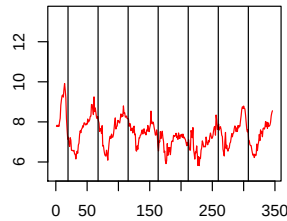
Taxa de concordância entre as duas partições: 82%

Análise de Clusters - k=3

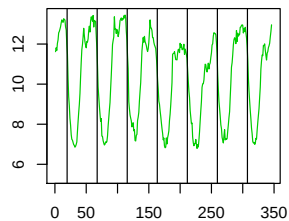
Completo



(a)

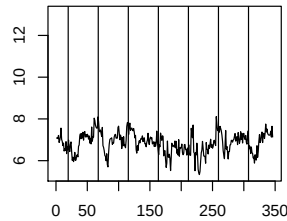


(b)

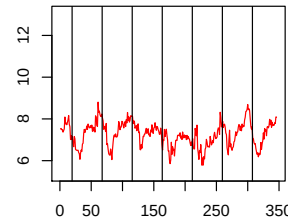


(c)

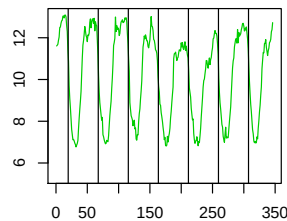
Truncado



(a)



(b)



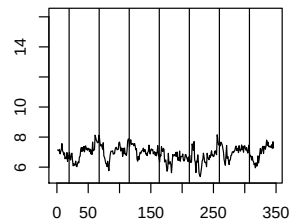
(c)

Taxa de concordância entre as duas partições: 82%

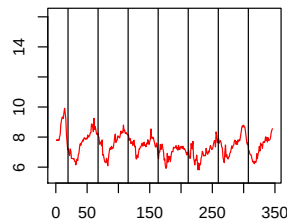
- **C1** - Taxas de transf. baixas ao longo de todo o dia.
- **C2** - Taxas de transf. baixas, mas mais elevadas que C1, especialmente no final do dia.
- **C3** - Taxas de transf. baixas no início e muito elevadas no final do dia.

Análise de Clusters - $k=4$

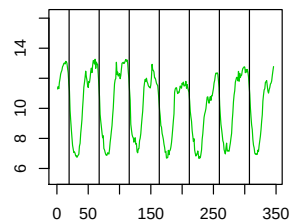
Completo



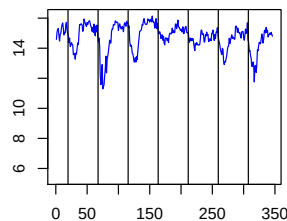
(a)



(b)

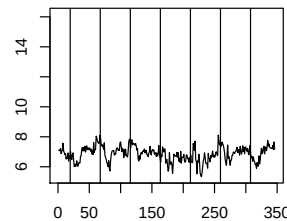


(c)

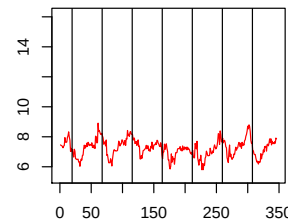


(d)

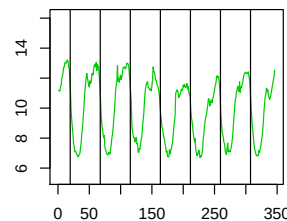
Truncado



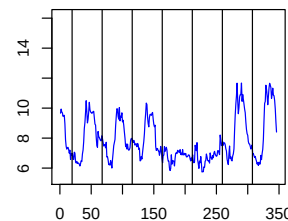
(a)



(b)



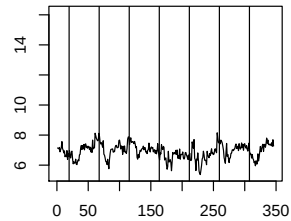
(c)



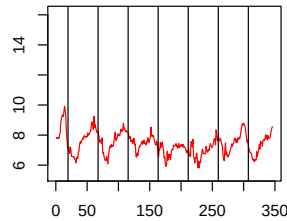
(d)

Análise de Clusters - $k=4$

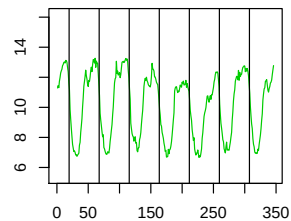
Completo



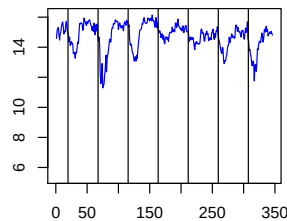
(a)



(b)

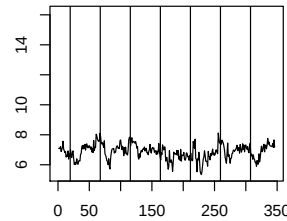


(c)

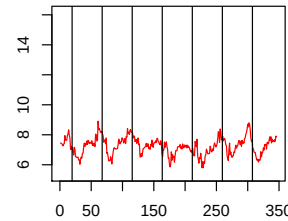


(d)

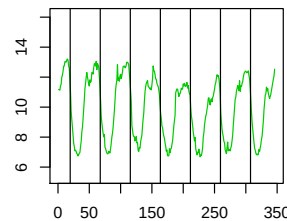
Truncado



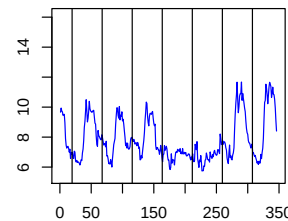
(a)



(b)



(c)

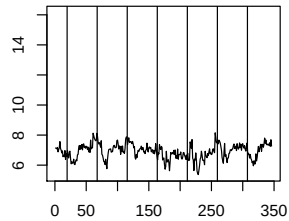


(d)

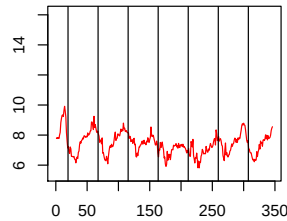
Taxa de concordância entre as duas partições: 51%

Análise de Clusters - $k=4$

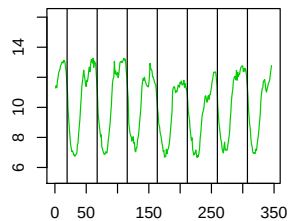
Completo



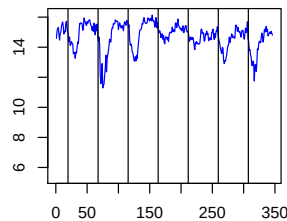
(a)



(b)

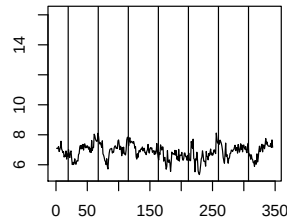


(c)

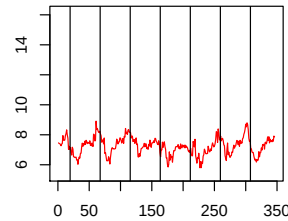


(d)

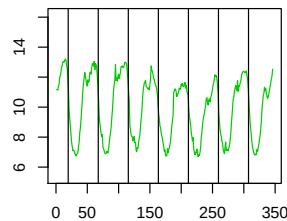
Truncado



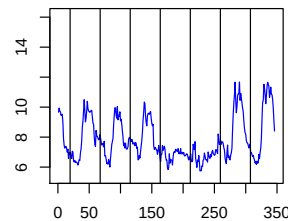
(a)



(b)



(c)



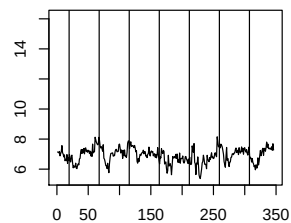
(d)

Taxa de concordância entre as duas partições: 51%

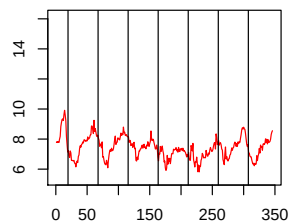
- **C1** - Taxas de transf. baixas ao longo de todo o dia.
- **C2** - Taxas de transf. baixas, mas mais elevadas que C1, especialmente no final do dia.
- **C3** - Taxas de transf. baixas no início e elevadas no final do dia.

Análise de Clusters - $k=4$

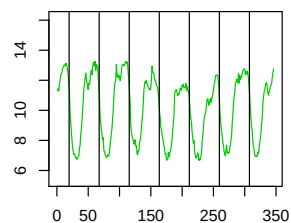
Completo



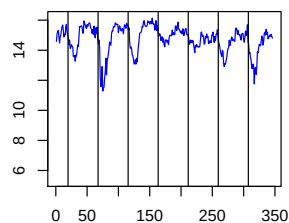
(a)



(b)

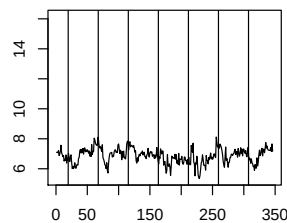


(c)

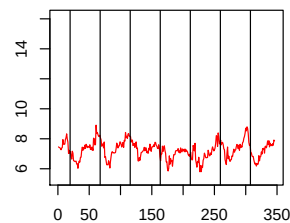


(d)

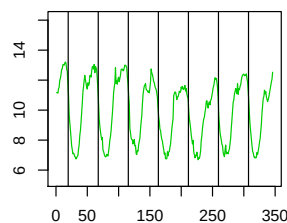
Truncado



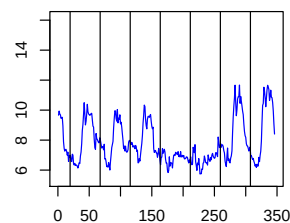
(a)



(b)



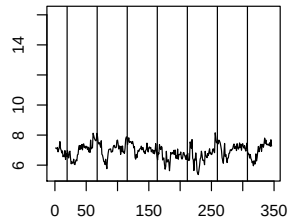
(c)



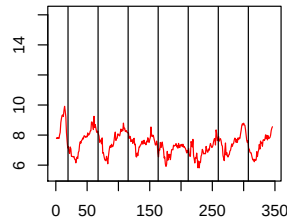
(d)

Análise de Clusters - $k=4$

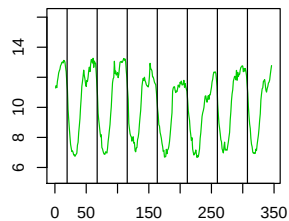
Completo



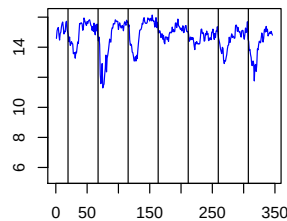
(a)



(b)

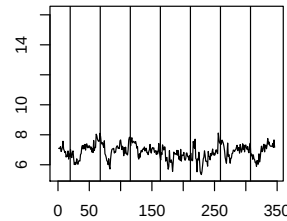


(c)

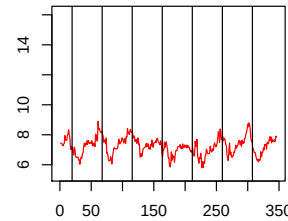


(d)

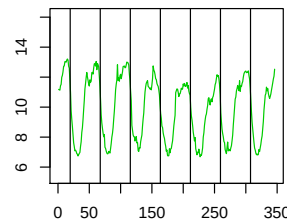
Truncado



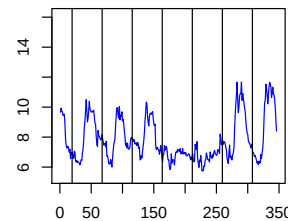
(a)



(b)



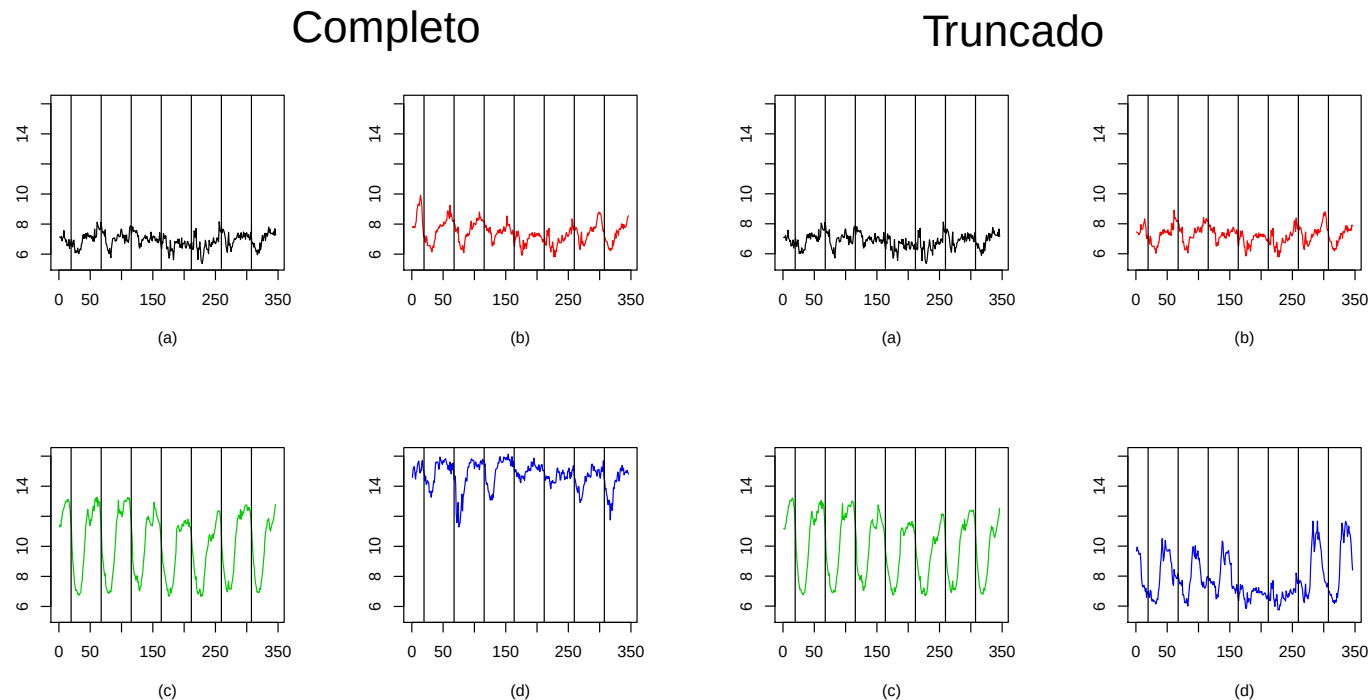
(c)



(d)

— **C4 - Completo** - Taxas muito elevadas ao longo de todo o dia.

Análise de Clusters - k=4



- **C4 - Completo** - Taxas muito elevadas ao longo de todo o dia.
- **C4 - Truncado** - Provavelmente, **empresas** uma vez que os seus utilizadores se caracterizam por taxas baixas durante o fim-de-semana e noite e elevadas durante as horas de expediente.

Conclusões II

- Cada vector resposta, associado a um utilizador, foi decomposto em 3 componentes: **periódica**, **ruído** e **picos**.

Conclusões II

- Cada vector resposta, associado a um utilizador, foi decomposto em 3 componentes: **periódica**, **ruído** e **picos**.
- Os fluxos próprios associados à componente periódica, ruído e picos explicam 52%, 47% e 1%, respectivamente.

Conclusões II

- Cada vector resposta, associado a um utilizador, foi decomposto em 3 componentes: **periódica**, **ruído** e **picos**.
- Os fluxos próprios associados à componente periódica, ruído e picos explicam 52%, 47% e 1%, respectivamente.
- Os **dados truncados** (considerando apenas a componente periódica) foram usados para detectar perfis semanais de utilização da Internet.

Conclusões II

- Cada vector resposta, associado a um utilizador, foi decomposto em 3 componentes: **periódica**, **ruído** e **picos**.
- Os fluxos próprios associados à componente periódica, ruído e picos explicam 52%, 47% e 1%, respectivamente.
- Os **dados truncados** (considerando apenas a componente periódica) foram usados para detectar perfis semanais de utilização da Internet.
- Os perfis detectados são essencialmente explicados pela **componente periódica**.

Conclusões II

- A partição obtida à custa dos dados truncados é caracterizada por valores mais elevados de *average silhouette*, o que indica uma **estrutura** em clusters **mais vincada**.

Conclusões II

- A partição obtida à custa dos dados truncados é caracterizada por valores mais elevados de *average silhouette*, o que indica uma **estrutura** em clusters **mais vincada**.
- Só se encontram discrepâncias assinaláveis entre as partições obtidas a partir dos dados completos e truncados quando se consideram **4 clusters**.

Conclusões II

- A partição obtida à custa dos dados truncados é caracterizada por valores mais elevados de *average silhouette*, o que indica uma **estrutura** em clusters **mais vincada**.
- Só se encontram discrepâncias assinaláveis entre as partições obtidas a partir dos dados completos e truncados quando se consideram **4 clusters**.
- Os dados truncados revelam um agrupamento dos utilizadores (**empresas**) que não é detectado pelos dados completos.

Conclusões II

- A partição obtida à custa dos dados truncados é caracterizada por valores mais elevados de *average silhouette*, o que indica uma **estrutura** em clusters **mais vincada**.
- Só se encontram discrepâncias assinaláveis entre as partições obtidas a partir dos dados completos e truncados quando se consideram **4 clusters**.
- Os dados truncados revelam um agrupamento dos utilizadores (**empresas**) que não é detectado pelos dados completos.
- Entender a estrutura do tráfego da Internet é importante para várias actividades de **engenharia de tráfego** e ainda para definir políticas adequadas de **tarifação**.