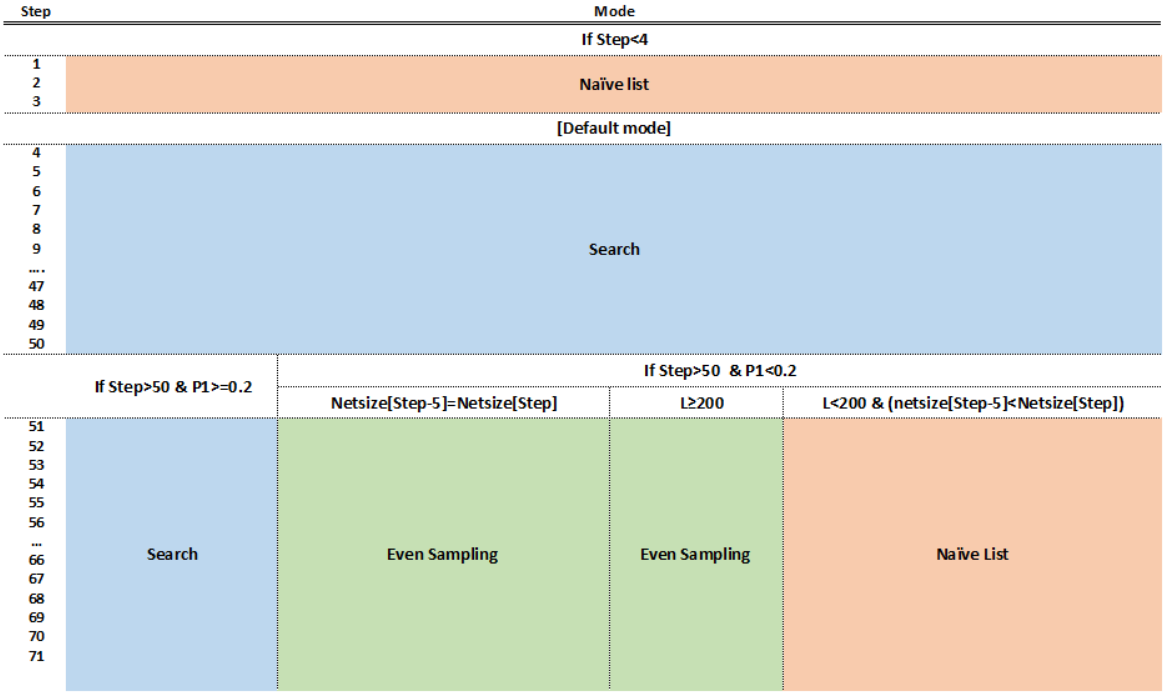
Network Sampling with Memory

Géraldine Charrance, Juillet 2019

1. Description de la méthode NSM

- Architecture générale

La méthode NSM (network sampling with memory) recourt à trois modes d'échantillonnage différents. Afin de déterminer le mode d'échantillonnage à mobiliser, on s'appuie sur une batterie d'indicateurs permettant de rendre compte du niveau d'exploration du réseau. Parmi eux, on trouve la taille du réseau dévoilé (L), le nombre d'interviews réalisées (Step) et la part de personnes citées une seule fois parmi toutes les personnes dévoilées (P1).



- Naive list mode

En début de collecte, on tire aléatoirement un individu parmi les premiers identifiés dans le réseau. Ayant une connaissance très faible du réseau à ce moment-là, il convient de mobiliser la technique de sondage la plus simple et égalitaire.

⇨ *Tirage aléatoire simple parmi tous les individus*

- Search mode

L'objectif est d'enquêter des pans du réseau encore inexplorés. Pour cela, il faut repérer, parmi les répondants, lesquels sont les plus susceptibles de nous y conduire. On calcule, pour chaque répondant, sa probabilité d'être un nœud-pont, basé sur la proportion de personnes citées une seule fois parmi leurs « amis ». Après identification des 5 individus les plus susceptibles d'être des nœuds-ponts, on en sélectionne un (proportionnellement à sa probabilité d'être un nœud-pont), puis on tire aléatoirement un de ses amis parmi ceux cités une seule fois et non enquêtés.

⇨ *Tirage à deux degrés : tirage d'un individu parmi les 5 individus les plus susceptibles d'être des nœuds-ponts, puis tirage d'un individu parmi les amis cités une fois et non interrogés du nœud-pont sélectionné.*

On mobilise ce mode après quelques steps afin d'orienter au mieux l'échantillonnage vers de nouveaux pans du réseau, et éviter de rester bloquer dans un cluster.

- Even sampling mode

L'objectif de ce mode est d'homogénéiser les CSR (cumulative sampling rate), c'est-à-dire d'exposer au tirage les individus nouveaux ou jamais soumis au tirage jusqu'alors. Pour cela, on exclut du tirage les nœuds les plus exposés au tirage précédemment, et ayant un CSR important.

> ⇨ *Tirage aléatoire simple parmi les nœuds ayant un CSR inférieur à l'ESR (even sampling rate) (ou les 100 minimum si le volume de nœuds remplissant la condition est inférieur à 100).*

On mobilise l'even sampling lorsque P1 passe en dessous du seuil A1. Cela signifie que peu d'individus dans le réseau n'ont été cités qu'une fois, c'est-à-dire que l'on a atteint un certain niveau d'exploration. Il convient alors de passer en even sampling afin de soumettre aux tirages les individus jusqu'alors ignorés, et de niveler les CSR. De même, on considère que si les 5 dernières interviews n'ont pas permis de découvrir de nouveaux individus, on arrive également à un niveau d'exploration suffisant qui autorise de sortir du mode Search pour passer en even sampling.

2. Coté programmation

a) Calcul des indicateurs et choix du mode d'échantillonnage

Afin de déterminer le mode d'échantillonnage à utiliser (parmi les différents modes présentés dans le schéma en partie 1), il faut calculer un certain nombre d'indicateurs :

- le nombre d'interviews réalisées (STEP)
- la taille du réseau dévoilé (L)
- la proportion de nœuds cités une seule fois et non enquêtés parmi tous les nœuds cités (P1)
- la taille du réseau dévoilé en ôtant les 5 dernières interviews réalisées (L_5)

Une fois ces indicateurs calculés, on détermine le mode d'échantillonnage à utiliser à partir des conditions suivantes (par défaut, le mode à utiliser est le search) :

- Search mode est le mode par défaut
- Si Step<4 ▶ Naive list mode
- Si P1<A1 et Step>50 et L=L_5 ▶ Even sampling mode
- Si P1< A1 et Step>50 et L≥200 ▶ Even sampling mode
- Si P1< A1 et Step>50 et L<200 et L>L_5 ▶ Naive list mode

b) Tirage d'un individu

- Naive list mode

Le tirage en naive list mode est un sondage aléatoire simple sur l'ensemble des nœuds dévoilés au moment du tirage. Les éligibles au tirage sont donc tous les nœuds cités, et la probabilité de tirage est de 1/L (taille du réseau dévoilé).

- Search mode

Le tirage en search mode se fait en deux étapes. Il s'agit d'un tirage à deux degrés.

1er degré : Pour déterminer les éligibles au 1er degré, il faut calculer, pour chaque répondant, sa probabilité d'être un nœud-pont. Pour cela, on réalise les calculs suivants pour chaque répondant :

- Nombre de personnes citées par répondant (d_j)
- Nombre de personnes citées une seule fois et non interrogée par répondant (c_j)
- $p(X \geq c_j | j \in L) \cong \sum_{i=c_j}^{d_j} \binom{d_j}{c_j} P1^{c_j}(1-P1)^{d_j-c_j}$

- Taille du réseau estimée : $\hat{G} = \frac{L}{(1-P1)}$ (valeur identique pour tous les répondants)
- Probabilité d'être nommé : $p_{nom_j} \cong 1 - (1 - \frac{d_j}{G})^{Step}$
- Probabilité d'être un nœud pont : $p_{bridge_j} = 1 - (p_{nom_j} \times p(X \geq c_j | j \in L))$

Les individus sont ensuite ordonnés par ordre décroissant sur la valeur de Pbridge, en excluant les enquêtés n'ayant aucun ami cité une seule fois et non interrogé (car aucun tirage de second degré ne pourra être réalisé dans ce cas).

On sélectionne les 5 premiers nœuds, c'est-à-dire les 5 ayant la plus forte proportion d'être des nœuds-ponts. On tire ensuite aléatoirement un individu parmi ces 5 selon un sondage proportionnel à la valeur de Pbridge.

2e degré : Parmi les amis cités une fois et non enquêtés du nœud-pont sélectionné, on tire un individu (selon un sondage aléatoire simple).

- Even sampling mode

Le tirage en EVEN SAMPLING cible les personnes ayant été peu exposées aux tirages successifs jusqu'alors. Un tirage aléatoire simple est fait parmi les nœuds ayant un CSR inférieur à l'ESR (even sampling rate, correspondant à la somme des SRS depuis le passage à l'even sampling). Si le volume de personnes répondant à ce critère est inférieur à 100, le tirage se fait parmi les 100 nœuds présentant les CSR les plus bas.

Tout comme pour le tirage du naive list mode, on ne tient pas compte du fait que les personnes aient déjà été échantillonnées (et interrogées). Il s'agit donc d'un tirage avec remise, contrairement au SEARCH MODE.

c) Actualisation des CSR

Le CSR est une valeur à actualiser et à suivre durant tout le processus d'échantillonnage : il est un moyen de tracer l'exposition des individus aux tirages successifs. Son suivi permet de déterminer les individus éligibles à l'Even Sampling, et sert *in fine* à calculer les poids.

Après chaque tirage, il faut donc incrémenter le CSR des individus y étant soumis.

- Naive list mode

En Naive list mode, tous les individus dévoilés sont éligibles au tirage et ils ont tous la même probabilité d'être tiré au sort. Le CSR est incrémenté de la valeur suivante :

$$S\_Rate_{Naive\ list} = \frac{1}{L}$$

- Search mode

En search mode, le tirage se fait parmi les amis cités une fois et non interrogés de l'individu nœud-pont sélectionné (au 1er degré). On actualise donc le CSR de ces individus uniquement.

Le CSR, pour ces individus, est incrémenté par le taux de sondage suivant (correspondant au produit des probabilités de tirage à chaque degré):

$$S\_Rate_{Search} = \frac{p_{bridge_j}}{\sum_{k=1}^{5} p_{bridge_k}} \times \frac{1}{c_j}$$

- Even sampling mode

L'actualisation du CSR se fait pour tous les nœuds candidats au sondage aléatoire simple de l'even sampling. Cependant, en fonction de la valeur du CSR au moment du tirage, le CSR n'est pas incrémenté de la même valeur.

Pour les individus nouveaux ou jamais soumis au tirage (c'est-à-dire présentant un CSR nul), le CSR est alors incrémenté de la valeur de l'ESR (somme des SRS depuis le passage à l'even sampling).

$$S\_Rate_{Even}New = \sum_{Step_{Even}}^{Step} \frac{1}{E_{step}}$$

$Step_{Even}$ étant le step auquel l'even sampling a été enclenché et E_s le nombre d'éligibles au tirage en even sampling (c'est-à-dire le maximum entre 100 et le nombre d'individus dont le CSR est inférieur ou égal à l'ESR).

A contrario, pour les individus présentant un CSR non nul, leur CSR est incrémenté de la valeur du SRS (c'est-à-dire 1/nombre d'éligibles au tirage).

$$S\_Rate_{Even}Old = \frac{1}{E_{step}}$$

d) Calcul des poids

Le plan de sondage reposant sur des tirages aléatoires successifs, la probabilité finale de sondage d'un individu correspond à la somme des probabilités de tirage à chaque étape (c'est-à-dire le CSR).

Le poids de sondage est ensuite obtenu en prenant l'inverse du produit du CSR et de p_{nom}, qui correspond à la probabilité de l'individu d'avoir été nommé (cf. tirage d'un individu selon le search mode).

$$W_j = \frac{1}{CSR_j \times p_{nom_j}}$$

3. Application : Simulations sur un réseau connu

a) Description du réseau

Des simulations ont été réalisées sur un réseau de 400 nœuds organisés en 4 clusters comptant 100 personnes chacun. Chaque cluster représente donc 25% de la population.

	Taille du roster	Ind. Cluster 1		Ind. Cluster 2		Ind. Cluster 3		Ind. Cluster 4	
Cluster 1	5	400	80,0%	32	6,4%	33	6,6%	35	7,0%
Cluster 2	10	32	3,2%	900	90,0%	34	3,4%	34	3,4%
Cluster 3	15	33	2,2%	34	2,3%	1400	93,3%	33	2,2%
Cluster 4	20	35	1,8%	32	1,6%	33	1,7%	1900	95,0%

Les réseaux personnels sont de taille et de structure différentes selon les clusters. Le cluster 1 se caractérise par des « rosters » de petite taille et davantage tournés vers les autres clusters.

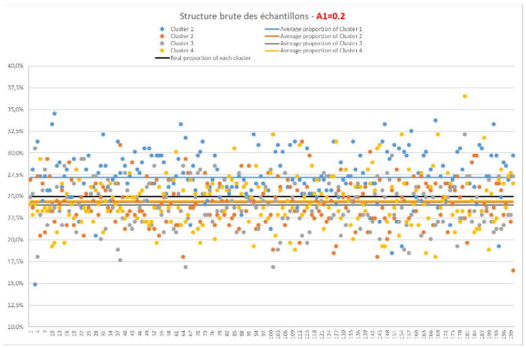
Pour vérifier le bon fonctionnement de l'algorithme programmé sous SAS, des simulations ont été réalisées (200 échantillons de 250 personnes).

b) Simulations comparées

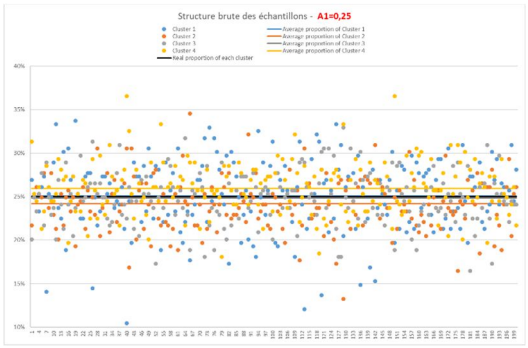
Dans un premier temps, nous avons conduit des simulations avec le seuil A1=0.2[1]. A chaque tirage, on calcule la valeur du paramètre P1 (proportion de nœuds cités une fois parmi tous les nœuds identifiés dans le réseau) et on la compare à A1. Si P1 passe en dessous de A1, on passe en even sampling. En augmentant la valeur de ce seuil, on sort donc plus tôt du mode Search et on passe donc plus rapidement à l'even sampling.

Lors de la première simulation avec A1 fixé à 0.2, les individus du cluster 1 étaient surreprésentés dans l'échantillon. Nous avons fait l'hypothèse qu'en augmentant le seuil à 0.25, on passerait plus vite en even sampling et on échantillonnerait moins de personnes en mode Search (mode qui semble surtout cibler les individus ayant des petits réseaux tournés vers l'extérieur). En passant plus en tôt en even sampling, on a plus de temps (de steps) pour rééquilibrer l'échantillon.

- <u>Structure brute des échantillons en termes de clusters</u>



Graphique 1a

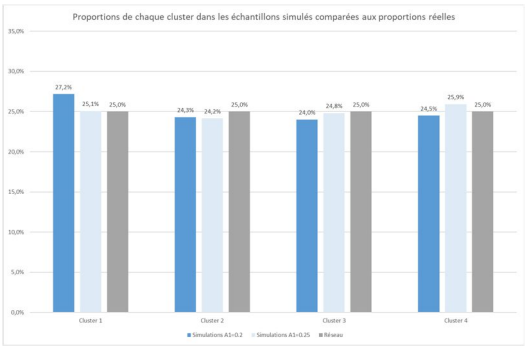


Graphique 1b

Simulations « A1=0.2 » : Par rapport à la structure réelle du réseau, les 200 échantillons présentent, en moyenne, une surreprésentation des individus du cluster 1 (27,2% de l'échantillon en moyenne contre 25% dans la population) et une légère sous-représentation des trois autres clusters (respectivement 24,3%, 24,0% et 24,5%).

Simulations « A1=0.25 » : La structure en fonction des clusters est assez proche de la structure réelle du réseau, bien que le cluster 4 soit un peu surreprésenté et le cluster 2 légèrement sous représenté. Les parts respectives des clusters sont en moyenne : 25.1% pour le 1er cluster, 24,2% pour le second, 24,8% pour le cluster 3 et 25,9% pour le cluster 4.

Comparaison des résultats : En comparant chacune des structures moyennes des échantillons avec la structure réelle, on observe de meilleurs résultats les simulations avec A1 fixé à 0.25. La distribution moyenne des clusters est plus proche de la réalité.

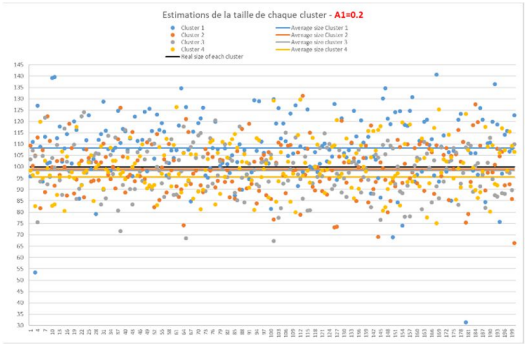


Graphique 2

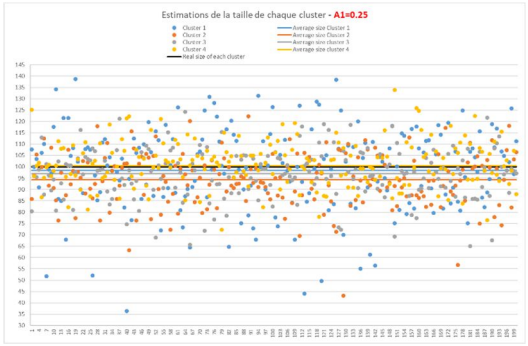
[1] Il faut noter que dans l'article de référence de 2012, A1 varie de 0.2 à 0.4, nous savions donc que nous pouvions tester différentes valeurs. Cependant, 0.4 nous semblait vraiment très élevé.

- **Tailles estimées des clusters**

On estime la taille de chaque cluster en faisant la somme des poids de sondage des individus par cluster.



Graphique 3a

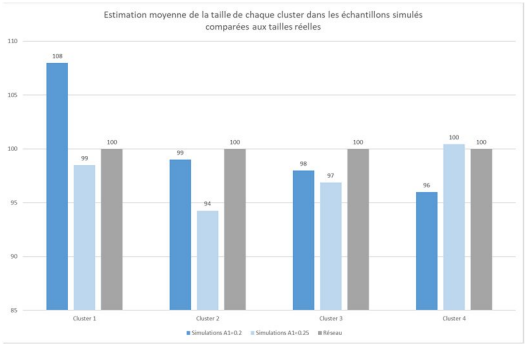


Graphique 3b

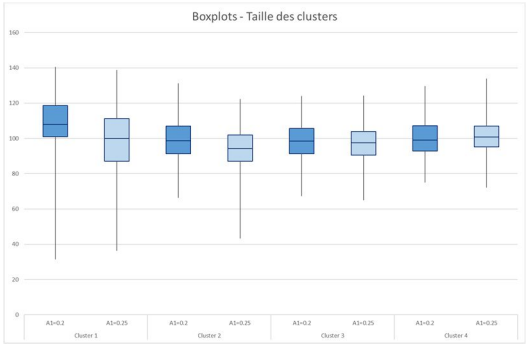
Simulations « A1=0.2 » : Tout comme précédemment, le cluster 1 se trouve surreprésenté en moyenne (taille moyenne estimée à 108 individus, le réseau comptant en réalité 100 personnes dans chaque cluster), et les trois autres clusters, surtout le 4, sont sous-représentés (estimations respectives : 99 personnes, 98 personnes et 96 personnes), lorsque la valeur de A1 est fixée à 0.2.

Simulations « A1=0.25 » : Par rapport aux simulations avec A1=0.2, on estime mieux la taille des clusters 1 et 4, et moins bien la taille des deux clusters intermédiaires.

Comparaison des résultats : Comme précédemment, en mobilisant la valeur du Khi², on conclut à une distribution plus proche de la distribution réelle avec A1=0.25.



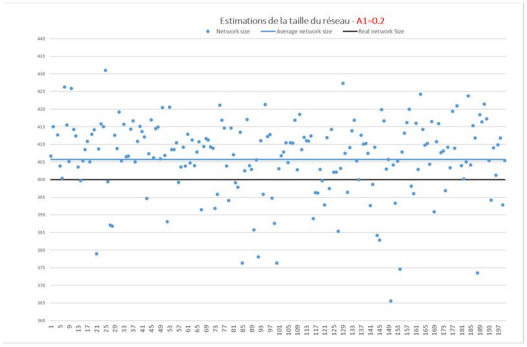
Graphique 4



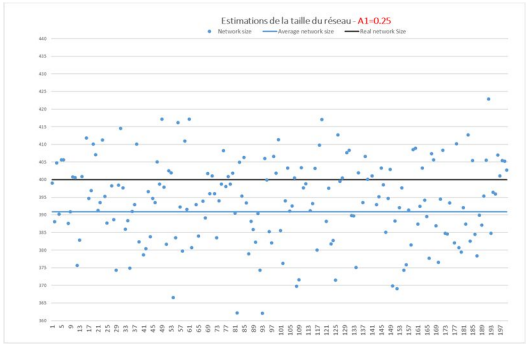
Graphique 5

- **Estimation de la taille du réseau**

L'estimation de la taille du réseau est obtenue en sommant les poids de sondage sur les 250 individus de l'échantillon.



Graphique 6a

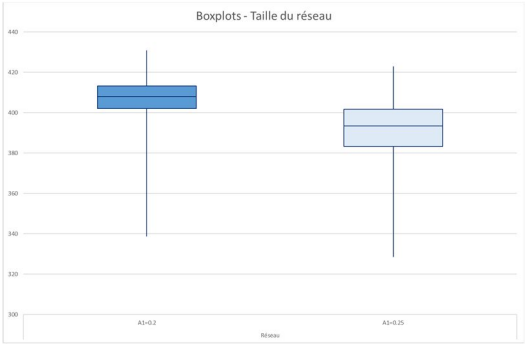


Graphique 6b

Simulations « A1=0.2 » : En moyenne on estime la taille du réseau à 406 individus. En fonction de l'échantillon, l'estimation de la taille du réseau varie entre 339 et 431, et la valeur médiane est de 408.

Simulations « A1=0.25 » : En moyenne on estime la taille du réseau à 391 individus. En fonction de l'échantillon, l'estimation de la taille du réseau varie entre 329 et 423, et la valeur médiane est de 393.

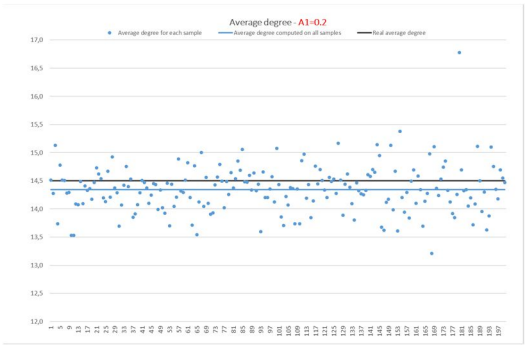
Comparaison des résultats : Dans le premier cas, on a tendance à surestimer la taille du réseau alors que dans le second, on la sous-estime. Cependant, l'avantage est donné, encore une fois, aux simulations avec A1=0.25 car la vraie valeur (n=400) est comprise dans l'intervalle interquartile (Q1-Q3).



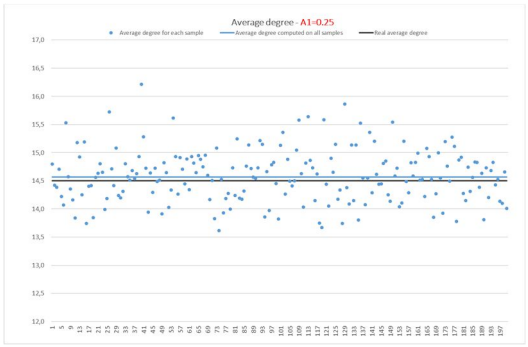
Graphique 7

- <u>Estimation de la taille moyenne des réseaux individuels (degree)</u>

A chaque cluster est associé une taille de réseau individuel (nombre d'amis) : 7 pour les individus du cluster 1, 12 pour le cluster 2, 17 pour le cluster 3 et 22 pour le cluster 4. En moyenne, sur le réseau complet, la moyenne de cette variable est de 14,5.



Graphique 8a

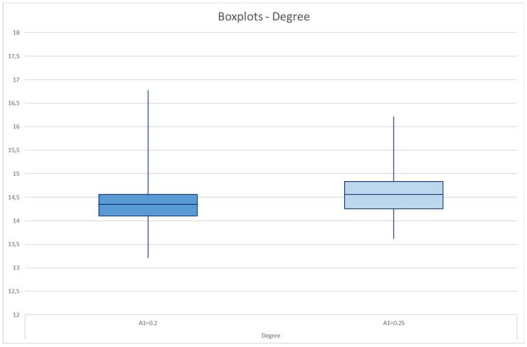


Graphique 8b

Simulations « A1=0.2 » : Les individus du cluster 1 étant surreprésentés en moyenne dans ces simulations, il en résulte une sous-estimation du « degree » moyen. En moyenne, on estime la taille du réseau individuel à 14,3. En fonction de l'échantillon, l'estimation varie de 13,2 à 16,8, avec une valeur médiane à 14,35.

Simulations « A1=0.25 » : En moyenne, on a tendance à surestimer la taille moyenne des réseaux individuels, 14,6 au lieu de 14,5. En fonction de l'échantillon, l'estimation varie de 13,6 à 16,2, avec une valeur médiane à 14,56.

Comparaison des résultats : Dans le cas des simulations avec A1=0.25, la moyenne et la médiane se trouvent plus proches de la vraie valeur que pour les simulations avec A1=0.2.



Graphique 9

- La structure des poids

Simulations « A1=0.2 » : La moyenne des poids est de 1,63 et varie de 0,21 à 1,88 en fonction des échantillons et des individus. La moyenne du coefficient de variation est de 10,77 et la moyenne des rapports max/min calculés pour chaque échantillon est de 3.98, ce qui signifie qu'en moyenne, l'individu « le plus lourd » pèse 3.98 plus lourd que l'individu « le plus léger ».

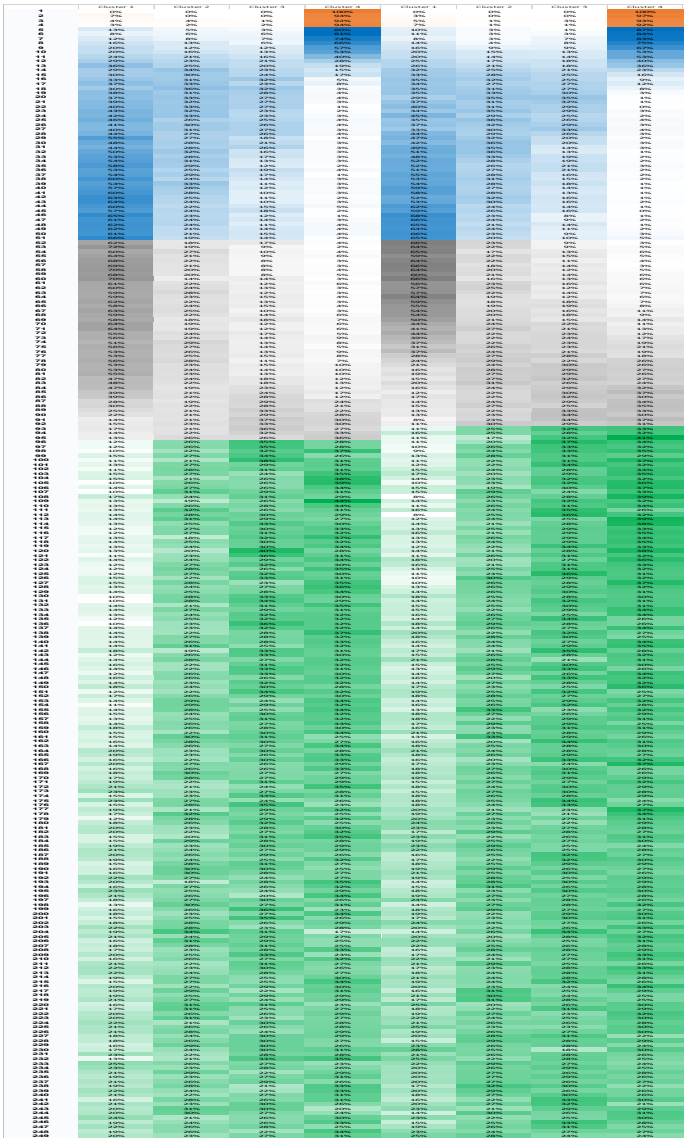
Simulations « A1=0.25 » : La moyenne des poids est de 1,57 et varie de 0,29 à 1,84 en fonction des échantillons et des individus. La moyenne du coefficient de variation est de 7,65 et la moyenne des rapports max/min calculés pour chaque échantillon est de 2,95.

Comparaison des résultats : Au vu des quelques statistiques de poids présentées ci-dessus, l'échantillonnage avec le paramètre A1=0,25 semble conduire à de meilleurs résultats, car moins de disparités de poids entre les individus donc des résultats plus stables.

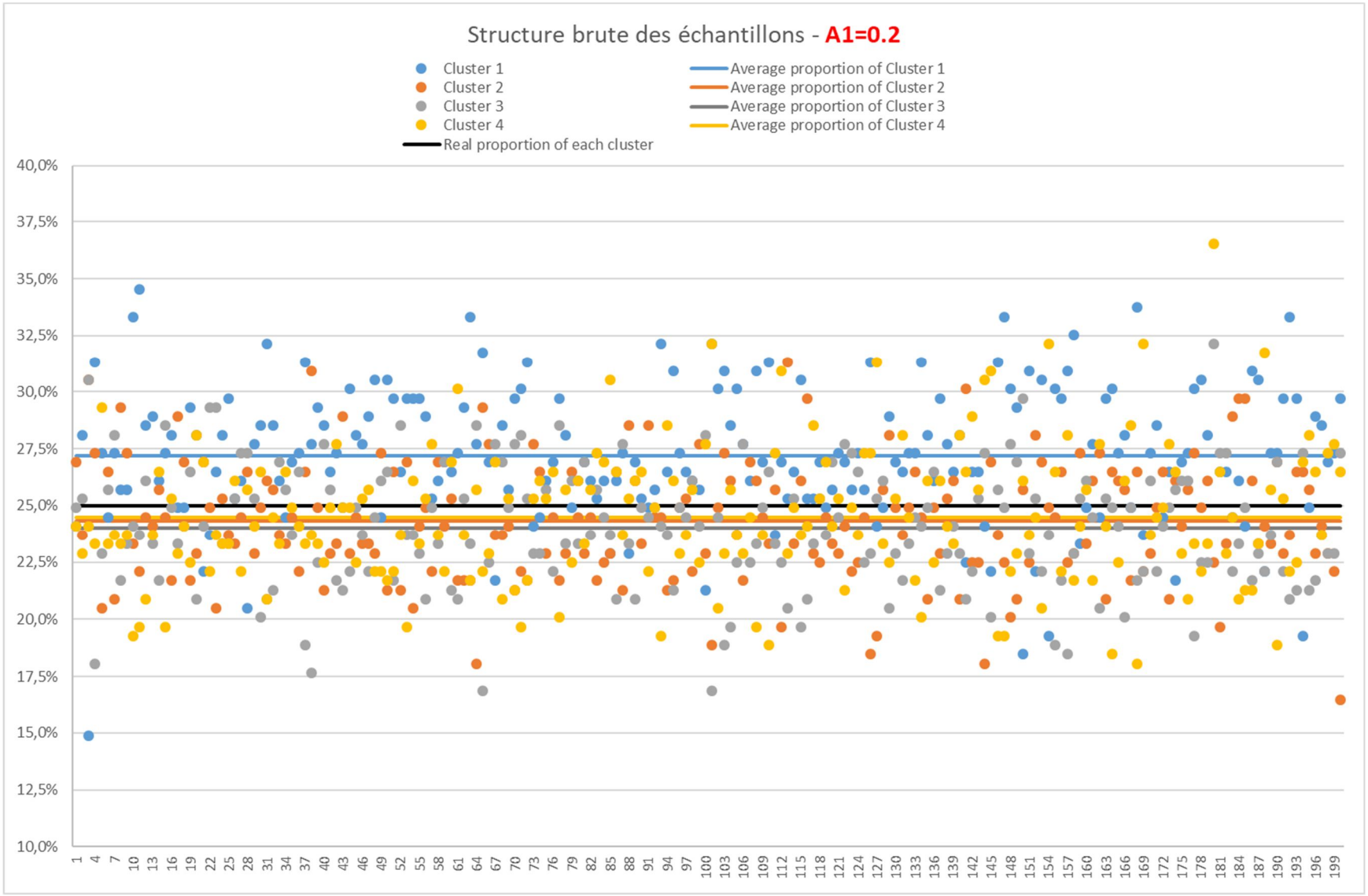
- Echantillonnage pas à pas

Les deux « tableaux » nous informe sur les clusters les plus échantillonnés à chaque step. Les steps en orange sont les steps où le mode de tirage est le NAIVE LIST MODE. La graine faisant partie du cluster 4 dans toutes les simulations, les premiers individus échantillonnés appartiennent également au cluster 4. Ensuite, sur les steps représentés en bleu, le tirage se fait selon le SEARCH MODE. Au départ, on échantillonne surtout des personnes du Cluster 4 car ce sont a priori, les seuls visibles à ce moment-là, puis, dès que des individus du cluster 1 sont identifiés, on se dirige vers ce cluster et on échantillonne en son sein[2]. La partie grise représente le passage entre le SEARCH MODE et l'EVEN SAMPLING. Ce changement s'opère entre le step 52 et le step 96 en fonction des échantillons. A partir du step 96, tous les échantillons emploient l'EVEN SAMPLING, on constate alors que le tirage se fait beaucoup moins au sein du cluster 1 que des autres clusters. C'est grâce à ces taux d'exposition différents aux différentes étapes qu'on obtient un échantillon équilibré du point de vue des clusters

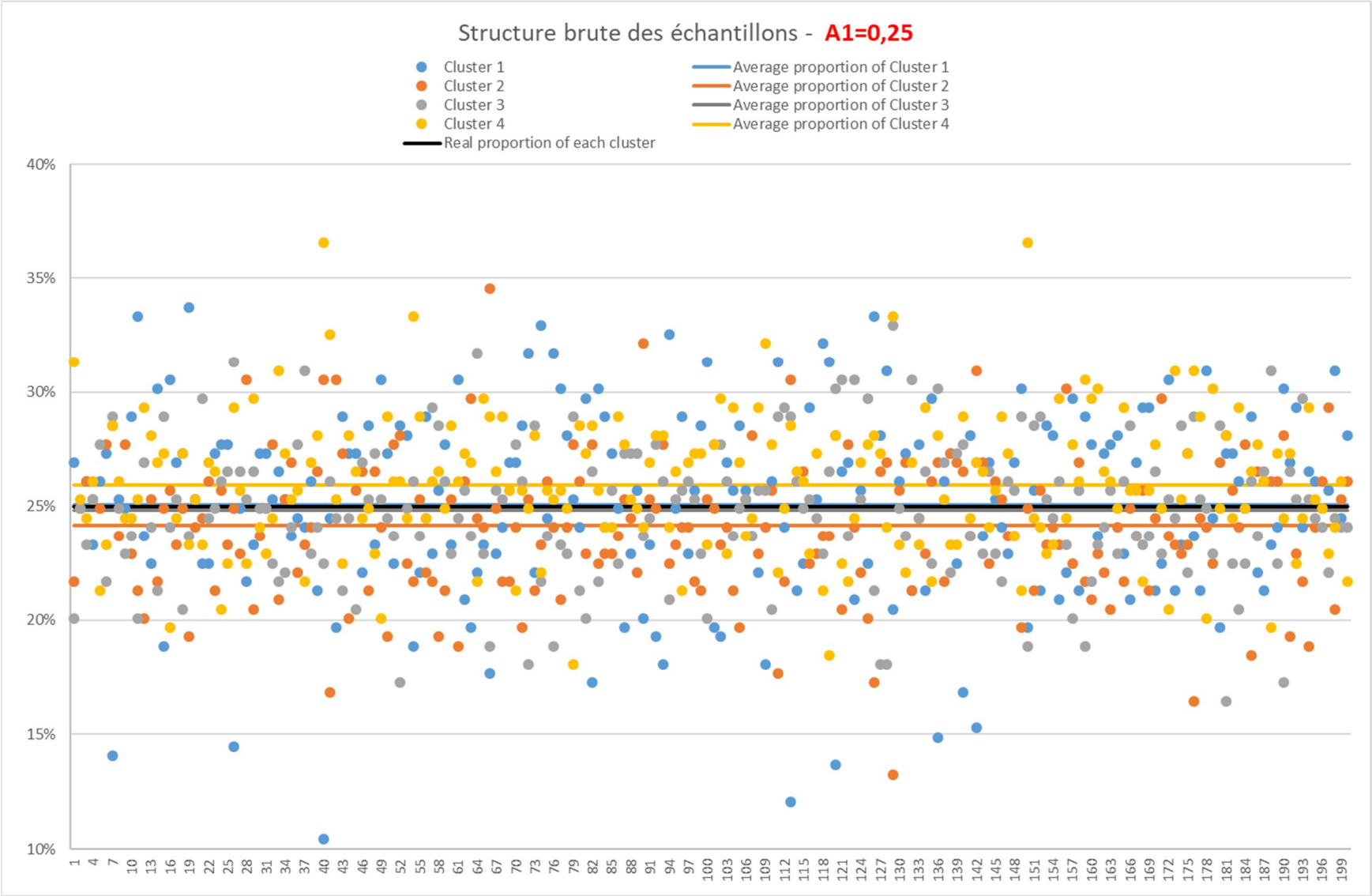
[2] Ce qui est exactement ce que l'on souhaite obtenir : le cluster 1 est plus ouvert donc contient a priori davantage de nœuds ponts vers d'autres clusters, c'est donc bien l'objectif du Search Mode d'identifier et de cibler ce cluster.



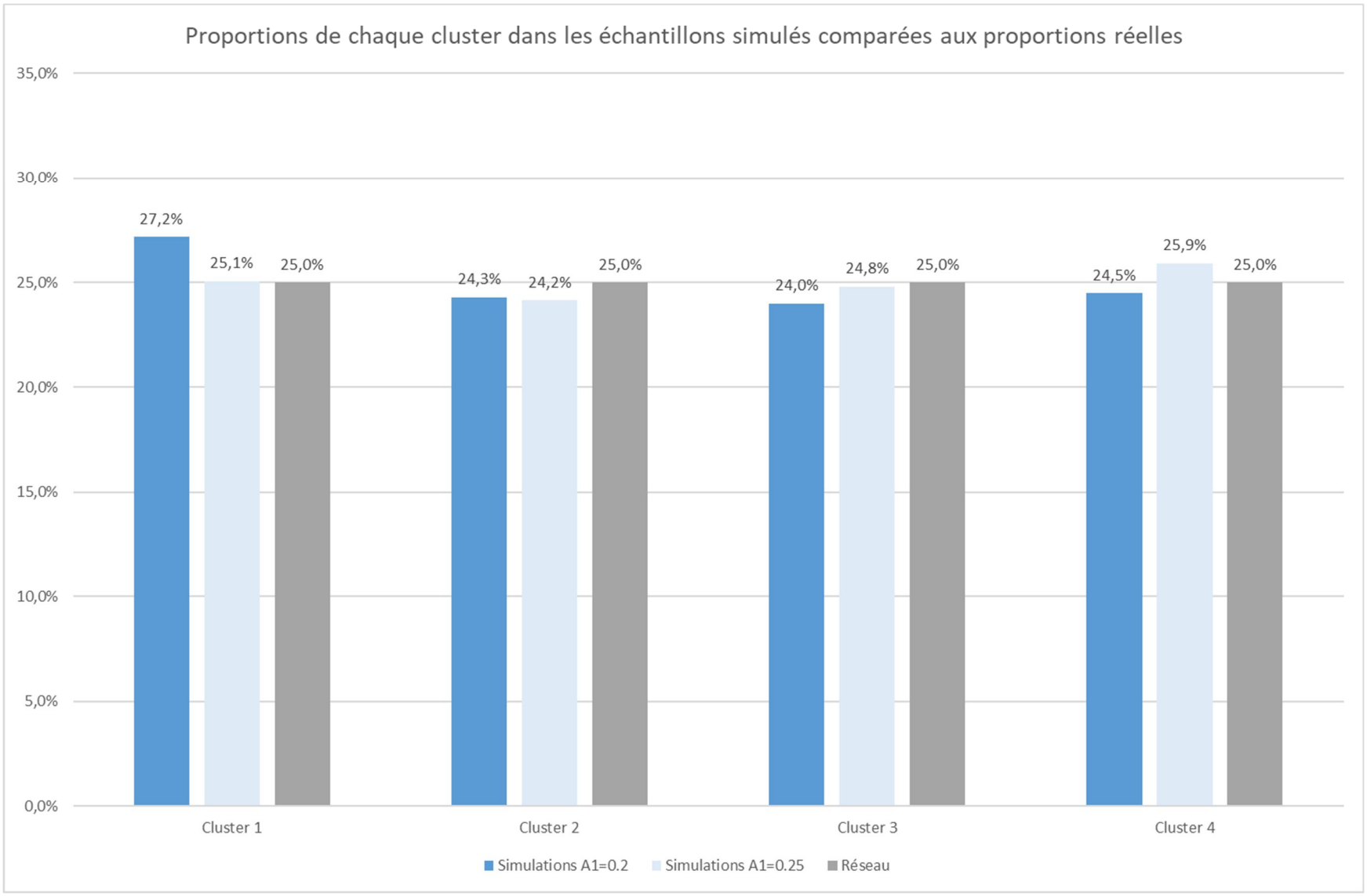
Graphique 1a



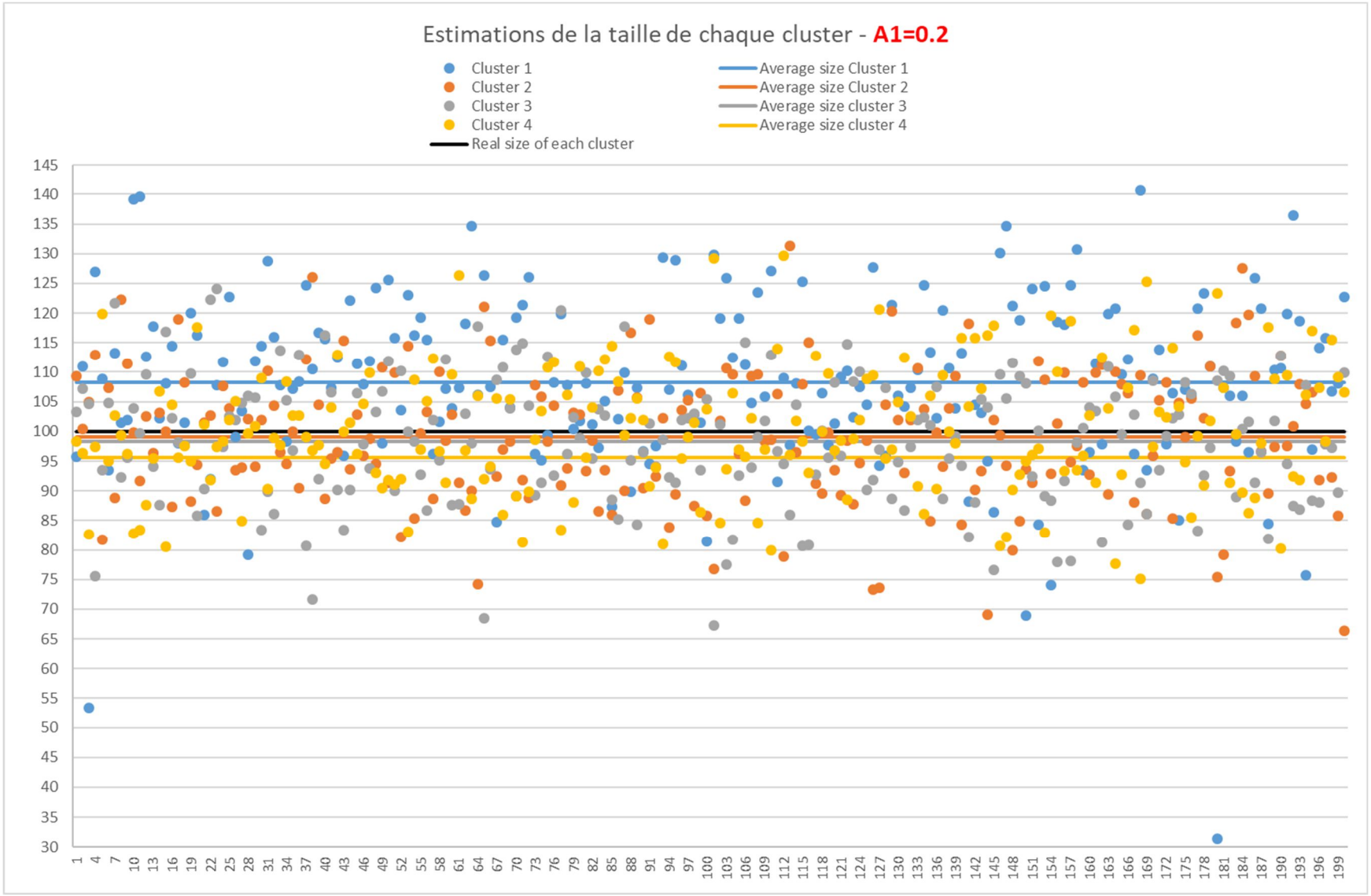
Graphique 1b



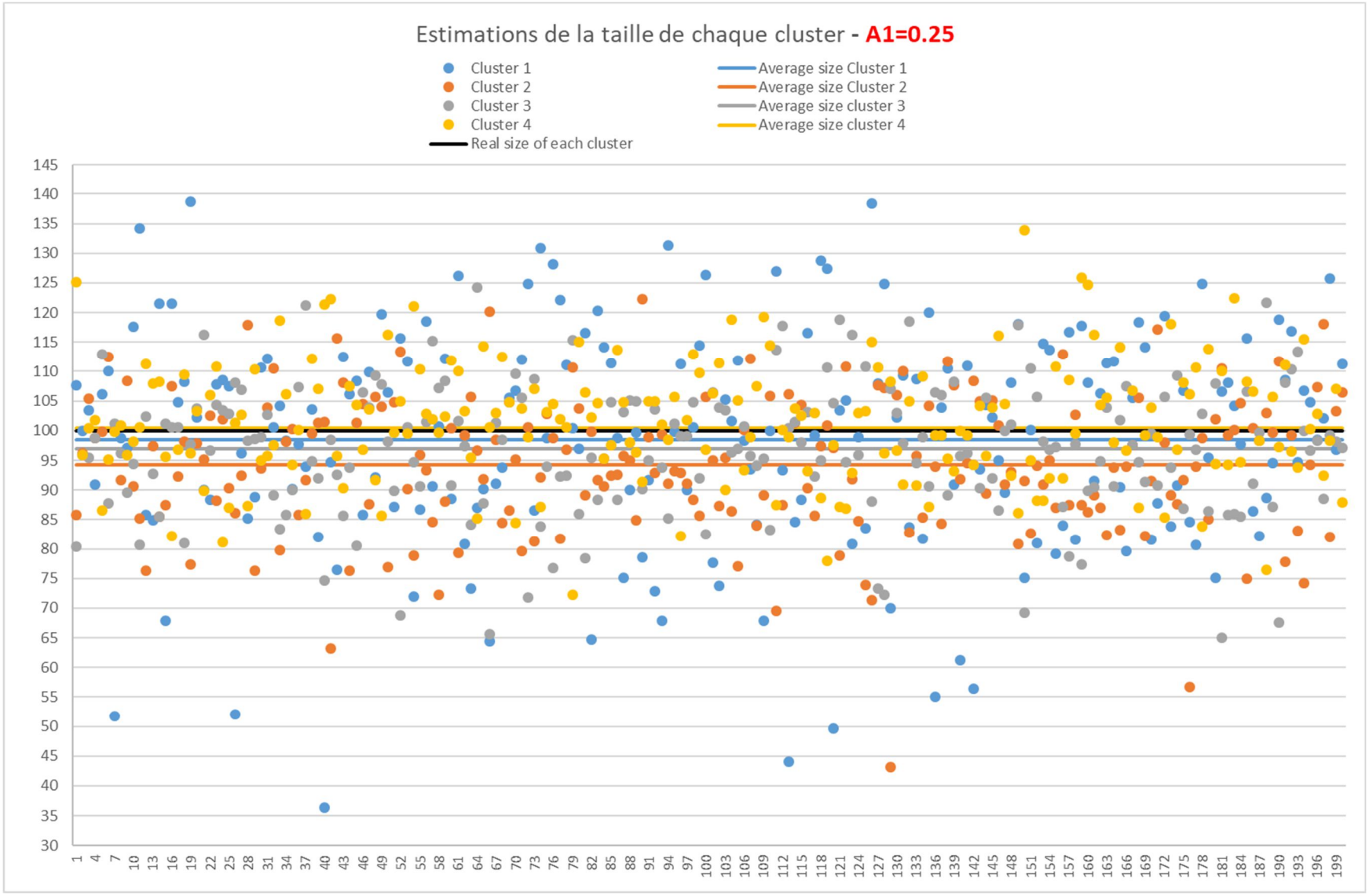
Graphique 2


Proportions de chaque cluster dans les échantillons simulés comparées aux proportions réelles
35,0%
30,0%
25,0%
20,0%
15,0%
10,0%
5,0%
0,0%
27,2%
25,1%
25,0%
24,3%
24,2%
25,0%
24,0%
24,8%
25,0%
24,5%
25,9%
25,0%
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Simulations A1=0.2
Simulations A1=0.25
Réseau

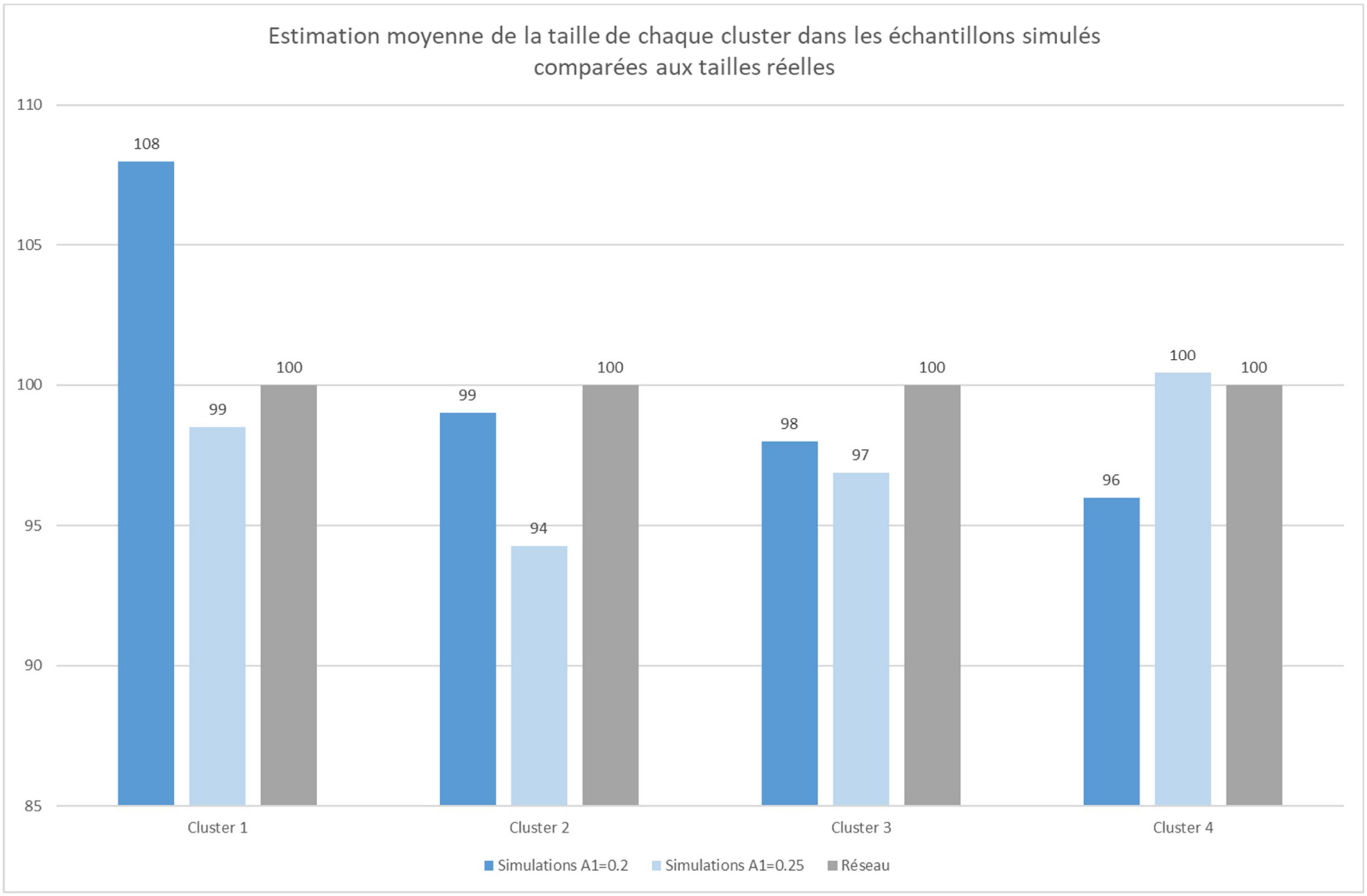
Graphique 3a



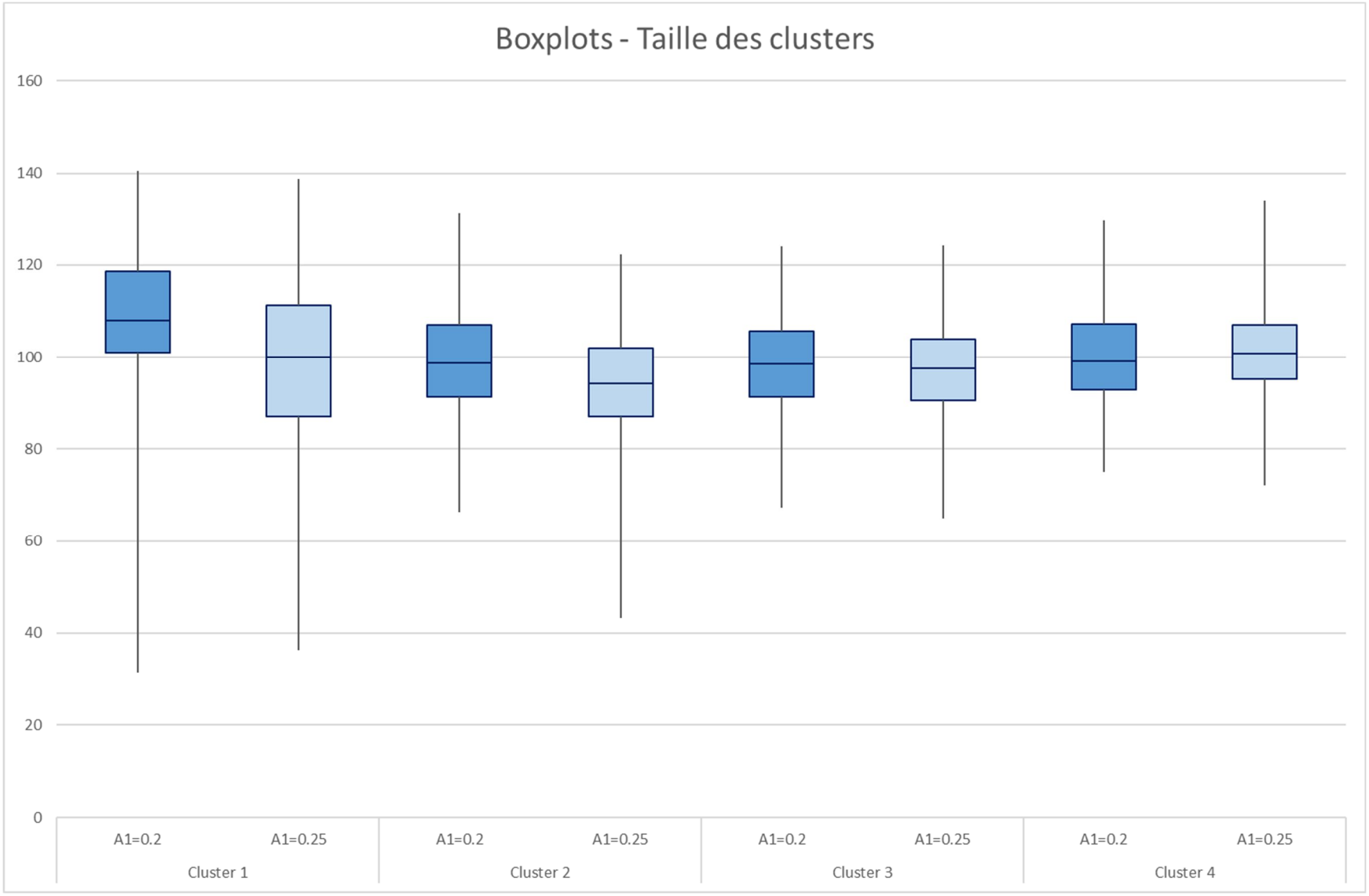
Graphique 3b



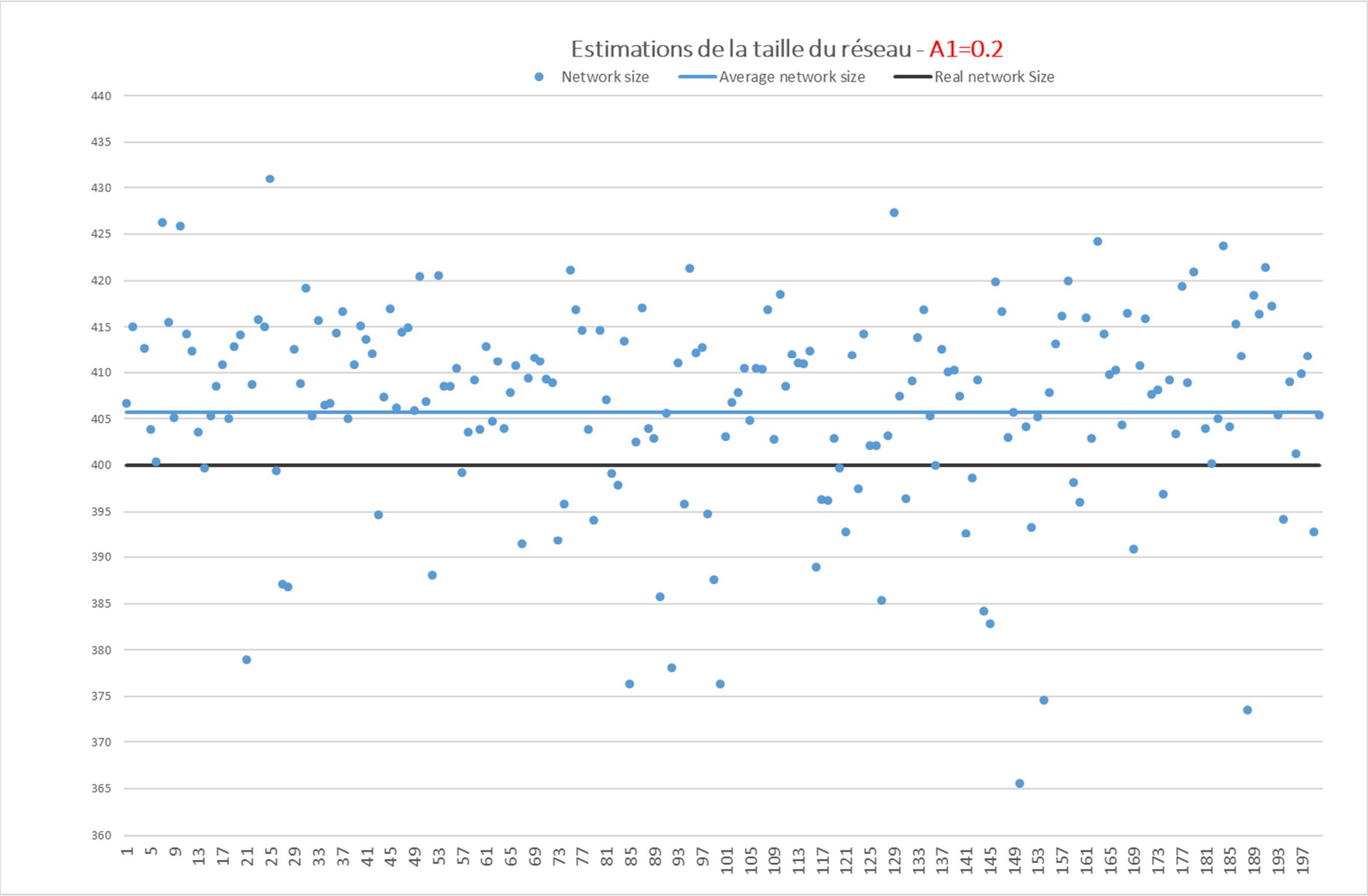
Graphique 4


Estimation moyenne de la taille de chaque cluster dans les échantillons simulés comparées aux tailles réelles
110
105
100
95
90
85
108
99
100
99
94
100
98
97
100
96
100
100
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Simulations A1=0.2
Simulations A1=0.25
Réseau

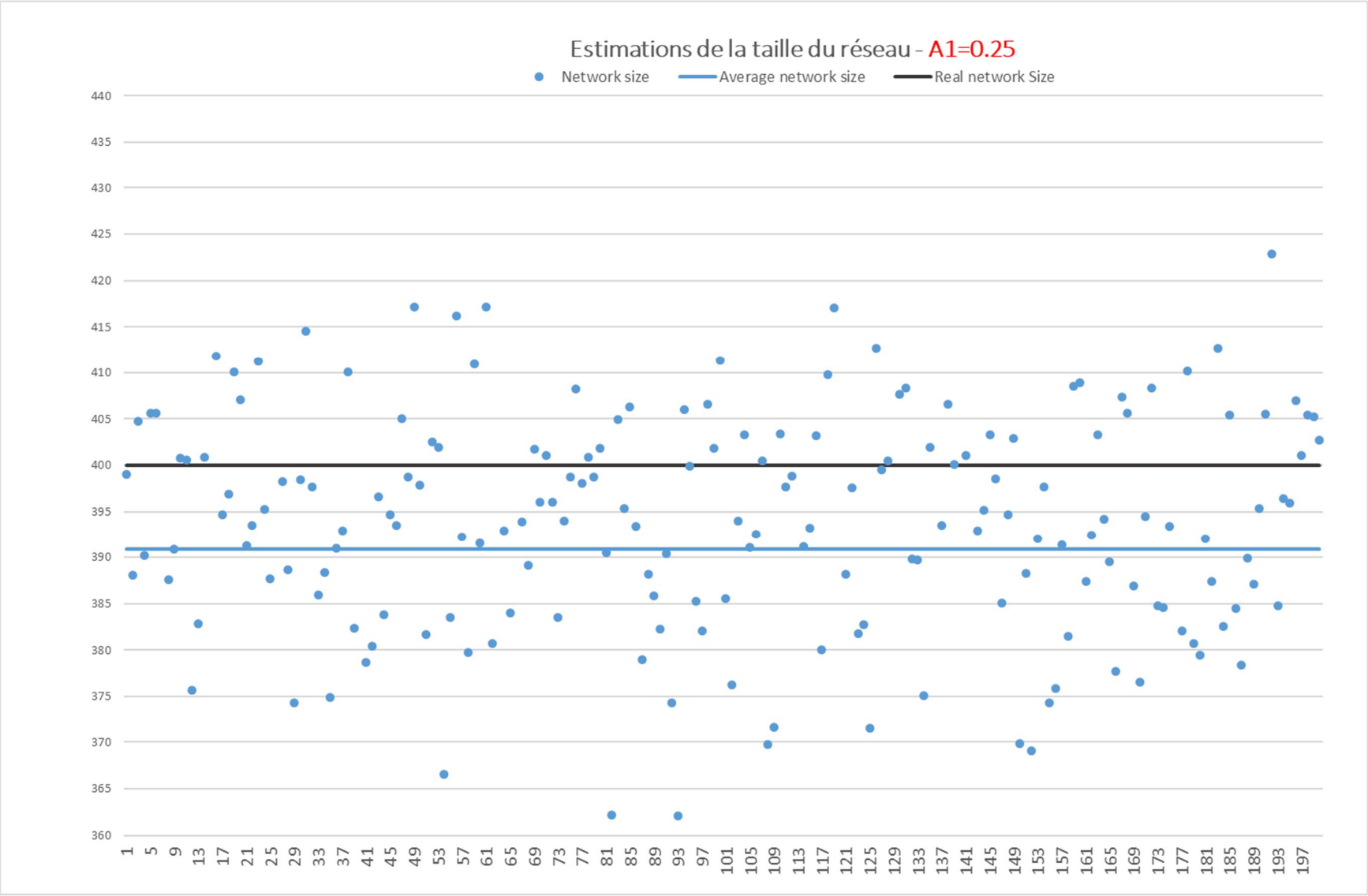
Graphique 5


Boxplots - Taille des clusters
160
140
120
100
80
60
40
20
0
A1=0.2
A1=0.25
Cluster 1
A1=0.2
A1=0.25
Cluster 2
A1=0.2
A1=0.25
Cluster 3
A1=0.2
A1=0.25
Cluster 4

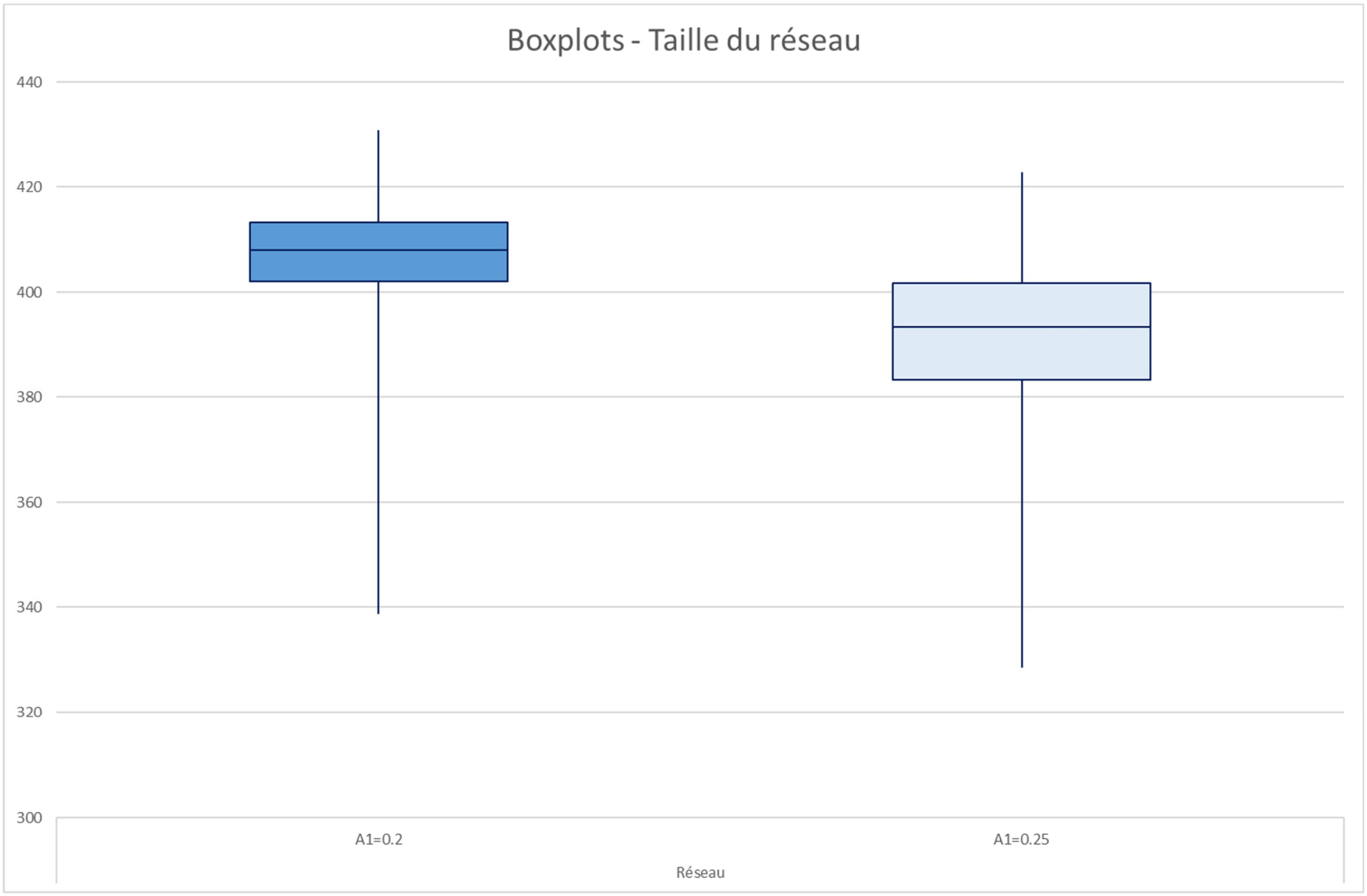
Graphique 6a



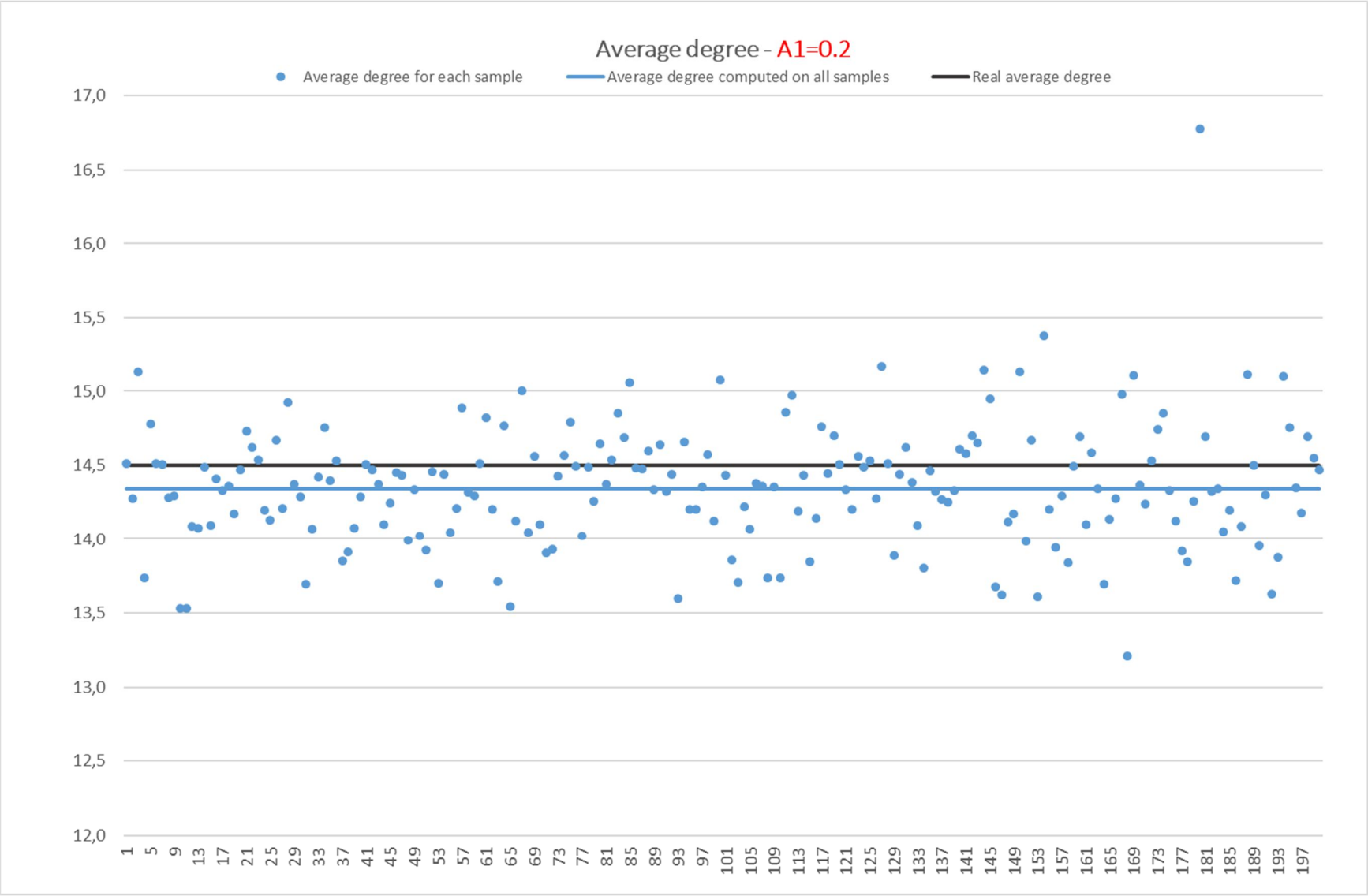
Graphique 6b



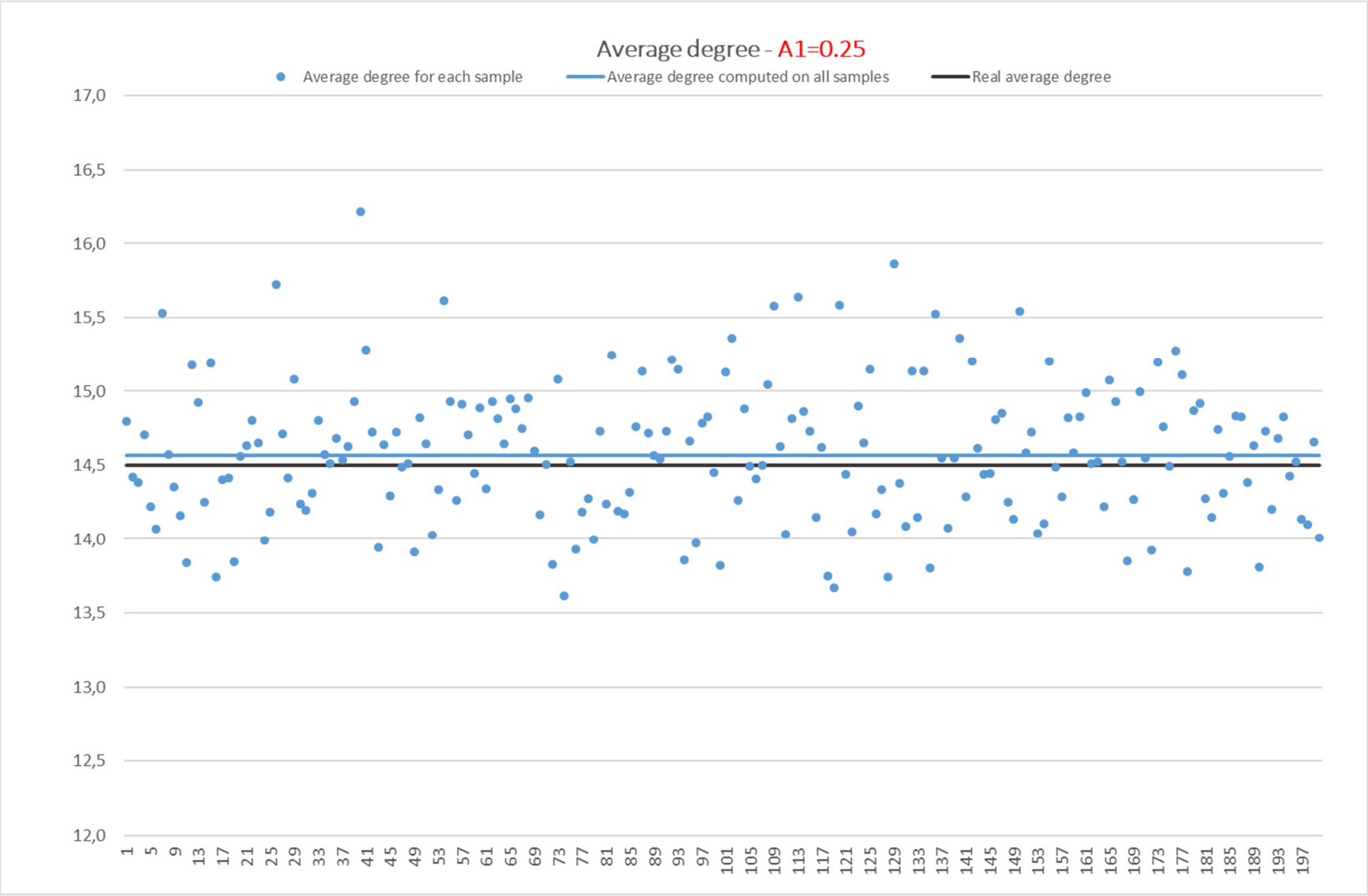
Graphique 7


Boxplots - Taille du réseau
440
420
400
380
360
340
320
300
A1=0.2
A1=0.25
Réseau

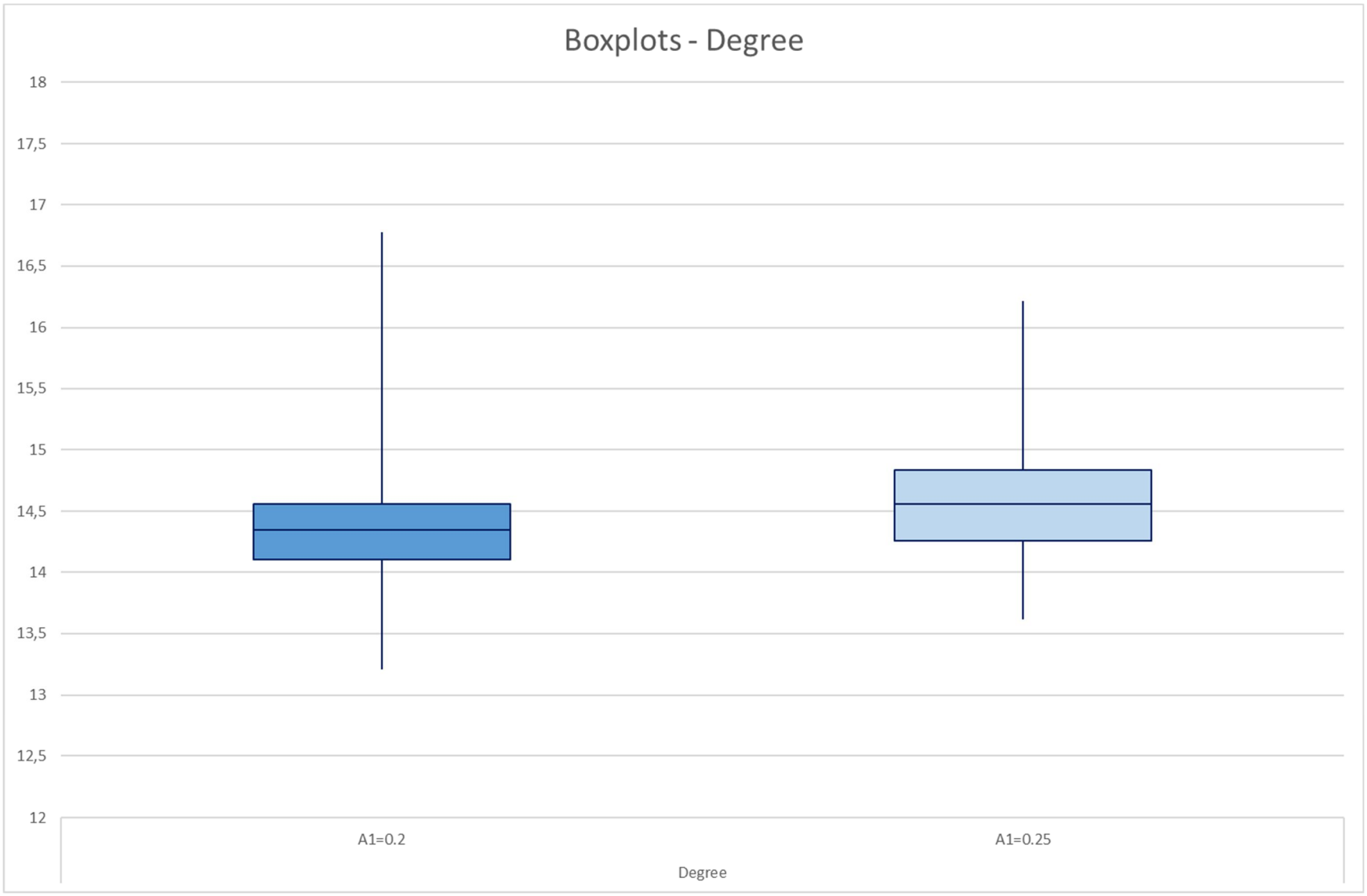
Graphique 8a



Graphique 8b



Graphique 9


Boxplots - Degree
18
17,5
17
16,5
16
15,5
15
14,5
14
13,5
13
12,5
12
A1=0.2
A1=0.25
Degree