Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

L'absence de données :

Une information précieuse

par Bernard Beauzamy

janvier 2019

1. Introduction

Cuvier pouvait, dit-on, reconstituer la forme d'un animal tout entier à partir de l'empreinte du pouce, c'est-à-dire produire un très grand nombre de données à partir d'une information très pauvre.

Le mathématicien peut faire mieux, ce qui illustre bien la supériorité des mathématiques sur la biologie. Il peut en effet reconstituer une quantité considérable d'information à partir d'une absence complète de données.

2. Trous dans les données

Imaginons qu'une population de plusieurs millions d'individus soit suivie ; on y rencontre des enfants comme des vieillards. On s'étonnera à bon droit si, au sein de cette population, on ne trouve personne dont l'âge soit entre 35 ans et demi et 36 ans. On se dira que la population est enregistrée à la naissance et que, pendant six mois, le système a été défectueux. La première idée qui vient, en pareil cas, est en effet celle d'une anomalie : chaque tranche d'âge devrait être représentée au sein de la population, à proportion de sa situation dans la pyramide des âges, qui est essentiellement continue, même en tenant compte des guerres.

Mais nous allons voir que cette idée de "continuité" est radicalement fausse ; il est normal que certaines séries de données comportent des "trous", dont la taille est tout à fait révélatrice de certains facteurs, liés à l'enregistrement.

3. Une anomalie apparente

Les données sur lesquelles nous travaillons ici sont des vitesses : vitesses de véhicules sur une bretelle de sortie d'autoroute, en l'occurrence l'autoroute A63 (Atlandes) ; nous les avons traitées initialement dans le cadre d'un contrat portant sur une autre question (détection des véhicules en contre-sens). Nous remercions Atlandes de nous les avoir communiquées.

Elles concernent plus de 180 000 véhicules ; un tiers au-dessus de 100 km/h ; deux tiers au-dessous de 90 km/h, mais aucune entre 90 et 100 km/h. A priori, il y a là une anomalie : il est impossible, sur une telle quantité de véhicules, que tous évitent soigneusement cette plage de vitesse.

En réalité, comme nous allons le voir, les données ne présentent aucune anomalie : le système d'enregistrement fonctionne correctement. Mieux, ce "trou" est porteur d'informations intéressantes.

4. Présentation des données

Les données consistent d'abord en des temps de passage sur deux boucles à détection magnétique, situées sur une bretelle de sortie de l'autoroute. Nous disposons de la différence $t_2 - t_1$ entre les temps de passage (en secondes) et, d'autre part, de la vitesse calculée par Atlandes (en mètres par seconde), pour chaque véhicule qui franchit les boucles. Il y a 182 661 données, recueillies sur une période de plusieurs mois. La plus grande vitesse est 28.86 m/s (soit environ 103.9 km/h) et la plus petite est 0.89 m/s (soit environ 3.2 km/h).

5. Analyse des temps

Voici le tableau des temps de passage enregistrés, arrondis à la seconde décimale par défaut :

Temps traversée (s)	nb occurrences	Temps traversée (s)	nb occurrences	Temps traversée (s)	nb occurrences
0,12	44559	0,72	28	1,23	2
0,14	74999	0,73	12	1,24	2
0,16	38516	0,74	12	1,26	3
0,18	15836	0,75	23	1,27	3
0,20	3979	0,76	22	1,28	1
0,22	1236	0,77	5	1,29	1
0,24	592	0,78	8	1,31	4
0,26	348	0,79	19	1,32	2
0,28	263	0,80	12	1,34	2
0,30	178	0,81	4	1,35	1
0,32	129	0,82	16	1,37	2
0,34	94	0,83	8	1,38	3
0,35	83	0,84	7	1,40	6
0,36	70	0,85	8	1,41	1
0,37	58	0,86	19	1,44	5
0,38	58	0,87	10	1,48	3
0,39	63	0,88	4	1,50	2
0,40	60	0,89	4	1,51	5
0,41	46	0,90	10	1,53	1
0,42	48	0,91	6	1,55	3
0,43	44	0,92	4	1,57	1
0,44	39	0,93	3	1,61	1
0,45	43	0,94	11	1,63	1
0,46	37	0,95	6	1,65	1
0,47	42	0,96	6	1,68	5
0,48	44	0,97	2	1,70	1
0,49	46	0,98	4	1,75	1
0,50	36	0,99	4	1,77	2
0,51	46	1,00	2	1,82	1
0,52	31	1,01	2	1,85	4
0,53	51	1,02	2	1,88	2
0,54	29	1,03	2	1,90	1
0,55	25	1,04	4	1,93	1
0,56	38	1,05	2	2,03	1
0,57	66	1,06	7	2,06	1
0,58	33	1,07	3	2,10	1
0,59	29	1,08	3	2,17	2
0,60	29	1,09	4	2,21	1
0,61	24	1,10	4	2,25	1
0,62	25	1,11	1	2,29	1
0,63	56	1,12	6	2,47	2
0,64	29	1,13	6	2,52	1
0,65	17	1,14	1	3,00	1
0,66	15	1,15	7	3,23	1
0,67	44	1,16	1	3,31	1
0,68	24	1,17	2	3,50	2
0,69	19	1,18	2	3,60	1
0,70	35	1,21	2	3,70	2
0,71	13	1,22	3	3,93	1

Tableau 1 : les temps de passage

On constate que pour les temps faibles, la précision est de 0.02 seconde ; elle passe à 0.01 seconde à partir de 0.35 s. Ceci est très commun et représente ce qu'on appelle un "effet d'échelle" (voir le livre [MPPR]) : le capteur est moins précis aux extrémités de la gamme de mesure. Pour les valeurs élevées, nous ne savons pas, faute d'enregistrement : le temps le plus élevé est 3.93 seconde et il n'y en a qu'un ; nous ne pouvons donc pas estimer la précision du capteur.

Nous remarquons ensuite qu'il y a très peu de "trous" dans les enregistrements de temps pour les valeurs faibles :

— Toutes les valeurs possibles sont représentées jusqu'à 1.18 seconde ;
— Il manque 1.19 et 1.20 et quelques-unes ensuite, jusqu'à 2.0 s ;
— Après cela, il y a peu d'enregistrements de temps supérieurs à 2.03 s.

Notre conclusion est donc simple : le capteur en temps fonctionne bien ; sa précision est meilleure que 0.02 seconde.

6. Analyse des vitesses

Les vitesses ne sont pas mesurées, mais calculées par Atlandes à partir des temps de passage, par la formule $v = \frac{d}{t}$, où d est la distance (connue de Atlandes) entre les deux boucles. Voici les relevés, par ordre croissant de vitesse.

v (m/s)	nb occ	v (m/s)	nb occ	v (m/s)	nb occ	v (m/s)	nb occ
0,89	1	2,78	3	4,14	7	6,78	46
0,94	2	2,81	2	4,19	8	6,97	36
0,97	1	2,83	2	4,22	10	7,06	46
1,00	2	2,86	3	4,25	6	7,25	44
1,06	1	2,89	2	4,28	4	7,36	42
1,08	1	2,94	2	4,33	6	7,58	37
1,17	1	2,97	2	4,36	6	7,69	43
1,39	1	3,00	1	4,39	7	7,92	39
1,42	2	3,03	7	4,42	12	8,03	44
1,53	1	3,06	1	4,44	1	8,31	48
1,56	1	3,08	6	4,47	7	8,42	46
1,58	1	3,11	6	4,50	5	8,69	6
1,61	2	3,14	1	4,56	12	8,72	54
1,67	1	3,17	4	4,58	10	8,86	63
1,69	1	3,19	4	4,64	9	9,17	58
1,72	1	3,22	3	4,67	14	9,33	58
1,81	1	3,25	3	4,72	12	9,67	70
1,83	1	3,28	7	4,75	12	9,86	83
1,86	2	3,31	1	4,81	12	10,25	94
1,89	4	3,33	1	4,83	16	10,89	129
1,92	1	3,36	4	4,89	13	11,58	1
1,97	2	3,39	2	4,94	18	11,61	177
2,00	1	3,42	2	4,97	17	12,42	263
2,06	1	3,44	2	5,03	19	13,36	36
2,08	5	3,47	2	5,08	16	13,39	312
2,11	1	3,53	4	5,14	8	14,47	578
2,14	1	3,56	4	5,17	19	14,50	14
2,17	1	3,58	2	5,22	25	15,78	305
2,22	1	3,61	2	5,28	15	15,81	931
2,25	3	3,64	4	5,33	17	17,36	3979
2,28	1	3,67	6	5,44	29	19,25	152
2,31	5	3,69	5	5,50	32	19,28	15680
2,33	2	3,72	6	5,56	24	19,31	4
2,36	3	3,75	3	5,61	25	21,64	1172
2,42	5	3,78	1	5,69	24	21,67	37333
2,47	1	3,81	3	5,81	29	21,69	11
2,50	6	3,83	6	5,86	27	24,69	1602
2,53	3	3,86	6	5,89	2	24,72	72459
2,56	2	3,89	4	6,00	33	24,75	937
2,58	1	3,92	4	6,08	25	24,78	1
2,61	2	3,97	4	6,14	41	28,78	43480
2,64	2	4,00	10	6,22	38	28,81	1062
2,67	4	4,03	8	6,31	25	28,83	15
2,69	1	4,06	11	6,44	29	28,86	2
2,72	1	4,08	2	6,53	51		
2,75	3	4,11	6	6,69	31		

Tableau 2 : les vitesses de passage

On constate en particulier qu'il n'y aucune mesure entre 24.78 et 28.78 m/s, soit entre 89,2 km/h et 103,6 km/h. A priori, ceci paraît incompréhensible : nous avons plus de 180 000 véhicules, dont plus de 40 000 ont dépassé 100 km/h ; comment se fait-il qu'aucun ne soit entre 90 et 100 km/h ? La première idée qui vient à l'esprit est que ceci est anormal : la mesure des vitesses n'est pas faite correctement.

En réalité, ces "trous" sont parfaitement normaux, comme nous allons le voir.

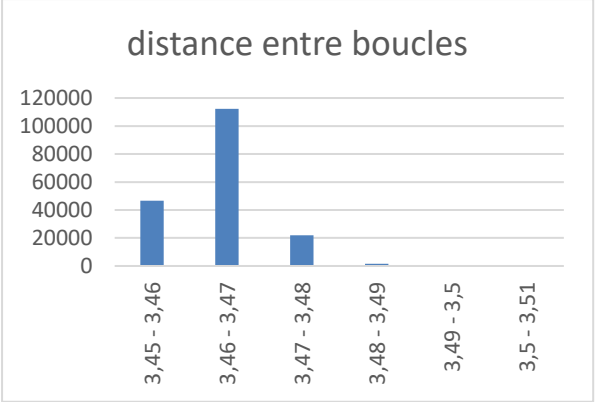
7. Calcul de la distance entre boucles

Intéressons-nous maintenant à la distance entre boucles. A partir des données ci-dessus, fournies par Atlandes (temps et vitesse), nous devons pouvoir reconstituer la distance entre boucles, par la formule $d = vt$. Mais on constate que les valeurs ainsi calculées ne sont pas toutes égales.

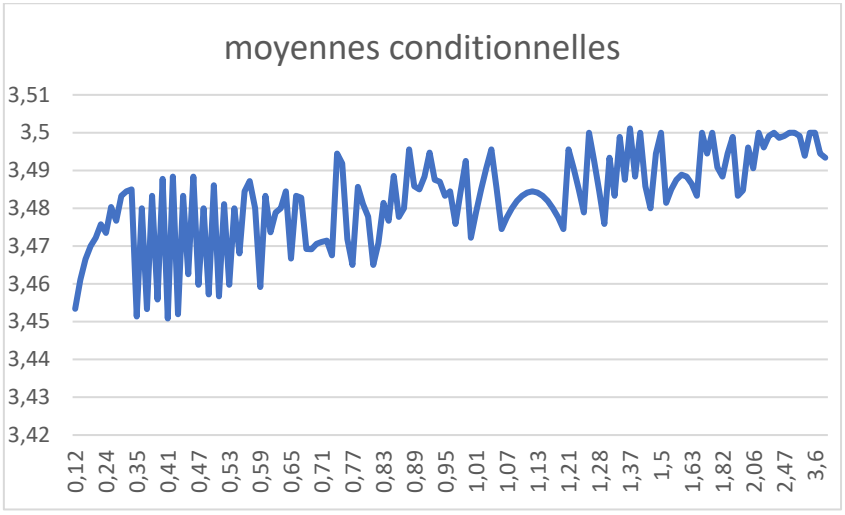
On peut d'abord se demander si cette distance est constante. Elle le serait évidemment s'il s'agissait d'une détection mécanique (les boucles sont fixes). Mais il s'agit d'une détection magnétique, et un camion sera détecté avant une motocyclette. On peut néanmoins penser que les deux boucles sont identiques, et que ce phénomène de décalage sera le même pour les deux boucles ; en d'autres termes, la différence de temps ne dépend pas du type de véhicule. Nous admettrons donc que cette distance doit être bien définie, et que les variations enregistrées proviennent d'erreurs d'arrondi sur le temps et la vitesse.

Voici les valeurs calculées et une présentation sous forme d'histogramme :

intervalle	nb de données
3,45 - 3,46	46666
3,46 - 3,47	112348
3,47 - 3,48	21912
3,48 - 3,49	1508
3,49 - 3,5	155
3,5 - 3,51	72



Il paraît légitime de prendre comme estimation de la distance entre boucles la moyenne des distances calculées sur l'ensemble des passages. Voici le graphe des distances calculées :



En abscisse, le temps de passage ; en ordonnée, la moyenne des distances calculées pour ce temps de passage. On constate que les valeurs augmentent en moyenne lorsque le temps augmente, mais on ne sait pas si le capteur est plus fiable pour des temps très importants. En conclusion, comme dit plus haut, il paraît raisonnable de prendre comme estimation de la distance entre boucles la moyenne de toutes les observations :

$$d_0 \approx 3.462 \text{ m}$$

8. Cas d'un chronomètre digital

Nous avons vu plus haut que, en regardant le tableau de données, on constate que le chronomètre a une précision d'au moins 2/100ème de seconde. Retrouvons ceci par le raisonnement, dans le cas d'un instrument "digital" (non analogique). Un tel instrument mesure une plus petite valeur (par exemple le centième de seconde) et toutes les valeurs mesurées sont des multiples de celle-ci.

Soit ε la plus petite valeur mesurée par le chronomètre. Toutes les valeurs seront de la forme $k\varepsilon$, k entier positif. On a $d = v_k k\varepsilon$ et donc $kv_k = \frac{d}{\varepsilon}$. Par conséquent, le produit kv_k doit être constant. On peut trouver la valeur de k associée à chaque valeur particulière de v. On écrit (en remarquant que v_k est une fonction décroissante de k) $kv_k = (k+1)v_{k+1}$, ou encore $k = \frac{v_{k+1}}{v_k - v_{k+1}}$.

Avec $v_k = 28.7\ m/s$ et $v_{k+1} = 24.7\ m/s$, on obtient :

$$k = \frac{24.7}{28.7 - 24.7} = \frac{24.7}{4} \approx 6$$

et enfin :

$$\varepsilon = \frac{d}{kv_k} \approx \frac{3.46}{6 \times 28.7} \approx 0.02\ s$$

autrement dit, le chronomètre a bien une précision de 2/100 de seconde.

9. Erreurs d'arrondi

A une valeur du temps unique (trajet entre les deux boucles), Atlandes n'associe pas une valeur unique de la vitesse, donc il y a des erreurs d'arrondi dans l'application de la formule $d = vt$ où d est fixe. Exemple : pour $t = 0.12$ s :

Temps traversée (s)	Vitesse véhicule (m/s)	Distance entre les boucles [m]
0,12	28,78	3,453
0,12	28,81	3,457
0,12	28,83	3,460
0,12	28,86	3,463

L'explication est simple : le temps réellement mesuré est variable, par exemple 0.1200, 0.1225, etc., mais le capteur tronque toujours à 0.12. Il fait le calcul de la vitesse avec le temps réel, affiche le temps tronqué et tronque la vitesse. Cela implique donc que le vrai temps est conservé et utilisé quelque part.

Dans le tableau des vitesses, on en trouve une à la valeur 24.78 m/s, puis aucune jusqu'à 28.78 m/s, qui est enregistré de très nombreuses fois. Cela correspond à des temps $t = 0.1396\ s$ pour la première et $t = 0.1202\ s$ pour les secondes. Toutes les vitesses entre les deux (elles sont très nombreuses) ont été reportées sur 28.78, c'est-à-dire sur $t = 0.12$ s.

En d'autres termes, l'absence de vitesses, dans le tableau des enregistrements, entre les valeurs 24.78 m/s jusqu'à 28.78 m/s, est simplement conséquence du fait que le chronomètre arrondit à 0.12 s tous les temps compris entre 0.12 et 0.14 s. Ce n'est en rien une anomalie.

Une même valeur de t, par exemple 0.12 s, peut donner quatre valeurs de la vitesse différentes, comme effet de la troncature. Prenons l'exemple des mesures de temps 0.1200, 0.1225, 0.1250, 0.1275, et calculons à chaque fois v par la formule $v = d / t$. On aura des valeurs de v différentes, à savoir 28.85, 28.26, 27.70, 27.15 m/s. Comme t est au dénominateur, plus t est petit et plus l'erreur sur la vitesse est grande.

Les deux boucles sont proches l'une de l'autre, parce qu'elles ne servent pas réellement à mesurer des vitesses, mais à détecter des passages en "contre-sens" (automobilistes prenant la bretelle de sortie à l'envers). Nous constatons que la précision de la mesure de vitesse est faible : inférieure à 10 km/h.

Pour obtenir une meilleure précision, il faudrait éloigner les deux boucles l'une de l'autre. Pour une précision de 1 km/h à 100 km/h, il faut que les boucles soient séparées d'au moins 55 m, mais alors elles n'évaluent plus qu'une vitesse moyenne. En outre, elles ne peuvent plus servir à la détection de contre-sens, parce qu'un véhicule peut faire demi-tour entre les deux.

Nous avons déjà obtenu cette conclusion dans le cadre d'un contrat avec la RATP : une vitesse ne se déduit pas d'une mesure du temps. La formule $d = vt$ est juste en théorie, difficile à appliquer en pratique.

Un conclusion complètement identique (trou naturel dans les données) apparaît si l'on cherche à appliquer la formule $F = m\gamma$ sous la forme $\gamma = \frac{F}{m}$. Imaginons le dispositif suivant : on cherche à déterminer l'accélération subie par un ensemble de billes, soumises à une force F constante et égale à $10\,N$. Les masses m sont mesurées à 0.2 gramme près et elles sont de l'ordre de 100 g. On ne trouvera aucune accélération dans la plage $\frac{10}{102} \le \gamma \le \frac{10}{100}$, c'est-à-dire entre 0.098 et $0.100\,m/s^2$.

10. Référence

[MPPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. SCM SA, ISBN 2-9521458-0-6. ISSN 1767-1175, broché, 369 pages. Seconde Edition, juin 2016.