ANALYSE DES COUTS DE DEPISTAGE DU SYSTEME D'INFORMATION VIBANQUE. (RETRIEVAL COSTS ANALYSIS OF THE VIBANK INFORMATION SYSTEM).

J.-E. Desgagnés, V. Srinivasan
Service d'analyse et d'indexation
Bibliothèque
Université Laval, Québec

RESUME

Description d'une procédure d'estimation des coûts de dépistage du système d'information automatisée VIBANQUE qui utilise pour le dépistage la programmation D.P.S. de I.B.M. (Describes the procedure for an evaluation of retrieval costs on the VIBANK automated information system utilizing I.B.M.'s D.P.S. programing).

INTRODUCTION

L'analyse des coûts d'un système d'information, qu'il s'agisse des coûts d'entrée, de stockage ou de dépistage de l'information, peut s'avérer très utile pour l'utilisation optimum des ressources disponibles lors de la planification budgétaire. Pour ce qui est des coûts d'entrée et de stockage de l'information, la détermination du coût par unité documentaire peut se faire assez facilement après une courte période d'opération par l'étude des dépenses affectées aux tâches d'analyse et d'indexation et de saisie des informations. Nous avons déjà effectué pour VIBANQUE une étude de ces coûts (SLINGERLAND, 1973). Il est par contre beaucoup plus difficile d'effectuer l'analyse des coûts de dépistage par ordinateur, coûts qui sont généralement fonctions de variables difficiles à identifier, ces variables étant en relation avec la programmation de dépistage utilisée. Le but de cette communication est d'établir une procédure d'estimation des coûts de dépistage du système d'information automatisée VIBANQUE, qui est en opération au Service d'analyse et d'indexation de la Bibliothèque de l'Université Laval et qui utilise pour le dépistage la programmation DPS (Document Processing System) de I.B.M.

CE QU'EST VIBANQUE

VIBANQUE est un système d'information spécialisée

sur la mécanique des vibrations et qui offre un service de recherche rétrospective dans ce domaine spécialisé. Une des particularités de VIBANQUE est que les résumés analytiques sont dépistés non seulement en utilisant les mots du titre, les noms d'auteurs et la date de publication du document mais en utilisant aussi près de 150 descripteurs génériques couvrant tous les sujets reliés à la mécanique des vibrations. La banque contenait au moment de notre étude, 17600 résumés analytiques et références bibliographiques et ce nombre s'accroit au rythme d'environ 5000 par année.

LE SYSTEME DPS

Le programme DPS que nous utilisons pour le dépistage est composé de six fichiers distincts. Pour VIBANQUE cependant nous n'utilisons que 4 de ces fichiers c'est-à-dire les fichiers "dictionnaire", "vocabulaire", "maître" et "texte", les deux autres, soit les fichiers "synonymes" et "équivalents", n'étant pas utilisés, ayant été jugés inutiles pour les besoins de VIBANQUE.

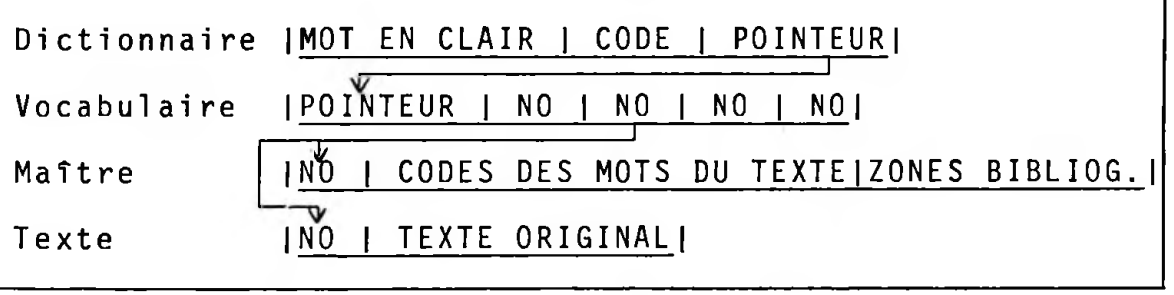
- Le fichier "dictionnaire" contient tous les mots en clair, suivis d'un code correspondant à chaque mot et d'un pointeur sur le fichier "vocabulaire". Dans la version originale du DPS, il y a, en plus de ce pointeur, deux autres pointeurs pour les fichiers "synonymes" et "équivalents".

- Le fichier "vocabulaire" contient chaque pointeur correspondant à un mot du "dictionnaire" suivi de la liste de numéros d'accession des documents contenant ce mot.

- Le fichier "maître" est composé de la liste des numéros d'accession des documents, chaque numéro étant suivi d'une part des codes des mots du texte de chaque document et d'autre part du contenu des zones bibliographiques. Ce fichier permet d'effectuer des recherches en tenant compte de la position des mots dans le texte comme par exemple une recherche sur les mots d'une même phrase, d'un même paragraphe.

- Le fichier "texte" est constitué des numéros de documents, suivis du texte original devant servir lors de la restitution des réponses.

FICHIERS DU DPS UTILISES
POUR VIBANQUE



CUEILLETTE ET ANALYSE DES DONNEES

Depuis le début des opérations de VIBANQUE, nous avons amassé, pour chaque requête ou chaque recherche effectuée dans la banque, les données qui pourraient éventuellement nous servir à élaborer une procédure nous permettant de prévoir les coûts d'une requête quelconque. Nous avons donc amassé les données concernant le temps CPU, le temps E/S (I/O), le Nombre de Références Dépistées (NRD), le nombre de références dans la banque (NRB) au moment de la requête et le coût de la requête. Nous avons dû abandonner l'idée d'établir une relation entre le coût et les variables micro (CPU, E/S, EXCP. etc.), étant donné que l'estimation de ces paramètres nous paraissaient fort complexe. Nous avons par contre mis en relation la variable NRD avec le coût de la requête et avons constaté que pour un Nombre donné de Références dans la Banque (NRB) le coût variait linéairement avec le NRD. (Voir figure 2).

$$Coût = aNRD + b \quad (1)$$

On voit également que la pente "a" qui représente le coût par référence dépistée, de même que la constante "b" qui représente le coût minimum d'une requête qui ne dépisterait aucune référence, varient légèrement avec l'augmentation du NRB. On peut voir à la figure 3 la variation de ces deux paramètres qui se traduisent par les équations suivantes:

$$a = \frac{-312.9\ NRB}{10^8} + \frac{905\ \sqrt{NRB}}{10^6} + \frac{1}{10^2} \quad (2)$$

$$b = \frac{2NRB}{10^5} + 2.696 \quad (3)$$

On peut expliquer la variation de ces paramètres de la façon suivante. Disons tout d'abord qu'au fur et à mesure que le nombre de termes dans le fichier "dictionnaire" augmente, le temps de recherche dans le fichier augmente également. Or étant donné que le type de requêtes que nous avons étudiées faisaient appel presqu'exclusivement au fichier "dictionnaire", on peut déduire que l'augmentation du "coût par référence dépistée" est due à l'augmentation du temps de recherche dans le "dictionnaire", les autres variables étant négligeables. Cependant, dans une banque d'information comme VIBANQUE où le vocabulaire est très spécialisé, le nombre de termes du dictionnaire tend à plafonner et à atteindre un maximum à mesure que l'on ajoute de nouvelle références. En effet, plus on ajoute de nouvelles références, plus l'augmentation proportionnelle du nombre de terme dans le dictionnaire est petite. Ceci explique le comportement de la variable "a" et le fait que le coût par unité dépistée augmente de moins en moins à mesure que le volume de la banque augmente et que le dictionnaire tend à atteindre son plafond.

Notre courbe de coûts cependant, ne serait probablement pas valable pour une autre banque d'information qui utiliserait pour le dépistage tous les fichiers du DPS ("Synonymes" et "Equivalents") de même que pour un type de questions où l'on ferait un plus grand usage du fichier maître. Cependant, il y aurait sûrement possibilité de déterminer empiriquement pour ce type de banque ou ce type de question, les courbes de coûts correspondantes.

Ceci étant dit, si l'on revient maintenant à nos formules et que l'on transpose les valeurs de "a" et de "b" dans l'équation (1) on a alors la formule qui nous permet de calculer le coût d'une requête en tenant compte à la fois du NRB et du NRD.

$$\text{Coût} = \left(\frac{-312.9\ NRB}{10^8} + \frac{905VNRB}{10^6} + \frac{1}{10^2}\right) NRD + \left(\frac{2NRB}{10^5} + 2.696\right) \quad (4)$$

Une telle formule n'aurait aucune utilité si on ne pouvait estimer le NRD avant d'effectuer une recherche quelconque dans la banque. Ceci nous est possible très facilement grâce à un outil que nous avons à notre disposition. Il s'agit d'une liste imprimée qui est un mélange du fichier "dictionnaire" et "vocabulaire" et dans laquelle nous avons par ordre alphabétique, les termes du dictionnaire, suivi de la liste des numéros d'accession des documents ou apparaissent chaque terme de même que leur fréquence d'apparition dans la banque. Lorsqu'une

requête nous est présentée, nous pouvons faire rapidement, à partir d'un échantillonnage, l'estimation du NRD. Nous pouvons aussi, avant d'effectuer une recherche, questionner la banque par télé-traitement et ne demander que les statistiques concernant le NRD. Dans le cas où le NRD nous semblerait excessif, nous modifions la requête jusqu'à ce que le NRD réponde à nos normes.

Si l'on analyse maintenant la table 2 dans laquelle on compare, pour les 25 requêtes qui ont servi à notre étude, le coût réel de la requête avec le coût tel qu'estimé par la formule, on peut voir que pour 22 données la marge d'erreur est de 5% et moins, pour 1 donnée, la marge d'erreur est de 7% et pour les deux autres données, la marge d'erreur est de l'ordre de 23% et 24%. Ces deux dernières données nous ont causé certains problèmes mais nous n'avons pu malheureusement trouver d'explication à leur comportement bizarre.

Pour continuer l'étude de la table 2, si l'on fait la somme des écarts en valeur réelle et que l'on divise le tout par la somme des coûts réels pour les 25 requêtes étudiées, on constate que la marge d'erreur globale n'est que de 2.4%.

CONCLUSION

Pour conclure, disons qu'une étude comme celle-ci peut s'avérer utile en ce sens qu'elle nous permet de faire des pronostics sur l'évolution des coûts de dépistage pour une banque en expansion continuelle (voir table 3). L'étude nous permet aussi de fixer pour chaque requête des usagers, un nombre maximum de références dépistées qui soit acceptable de même qu'un prix qui soit en accord avec le coût réel. L'idéal cependant serait d'effectuer une étude comparative de différents systèmes de dépistage, ce qui nous permettrait alors de juger de l'efficacité de ces systèmes.

REFERENCES BIBLIOGRAPHIQUES

SLINGERLAND, F.W., SRINIVASAN, V. "A mathematical model of an information system", dans American Society for Information Science, Proceedings of the 36th annual meeting, Los Angeles, oct. 21-25, 1973.

FIGURE 1

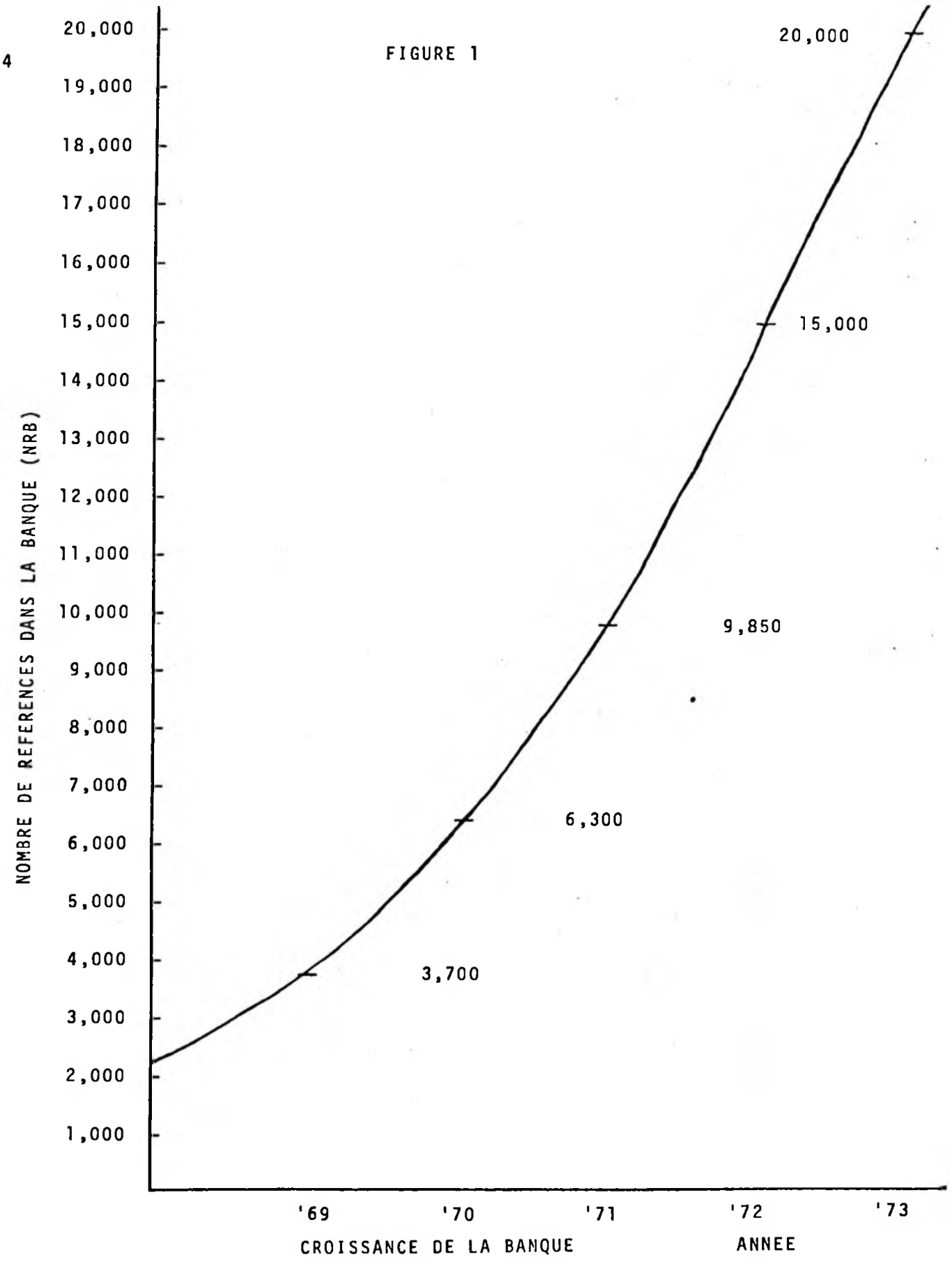
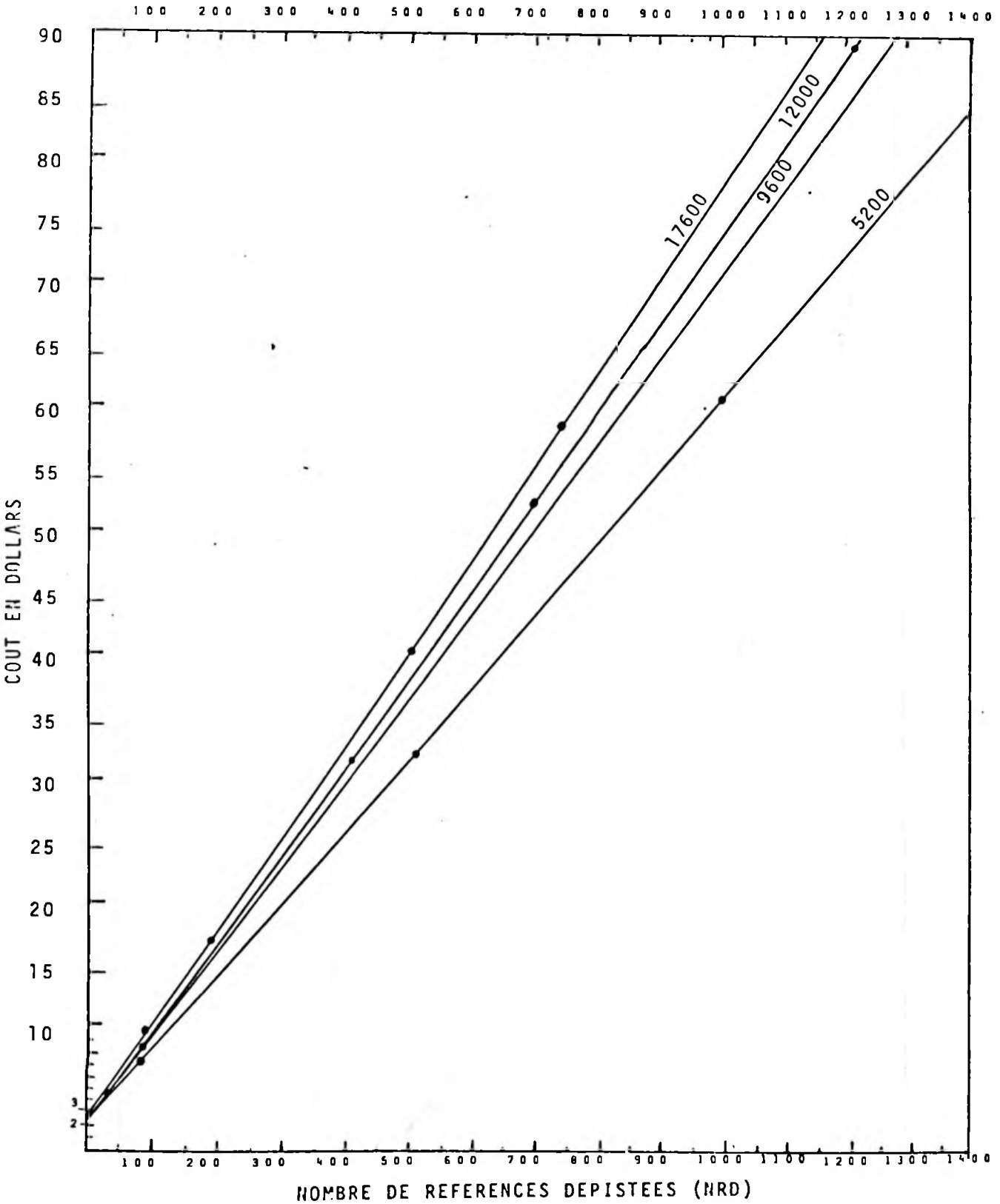

20,000
19,000
18,000
17,000
16,000
15,000
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
NOMBRE DE REFERENCES DANS LA BANQUE (NRB)
20,000
15,000
9,850
6,300
3,700
'69
'70
'71
'72
'73
CROISSANCE DE LA BANQUE
ANNEE

FIGURE 2 : COUT VS NRD



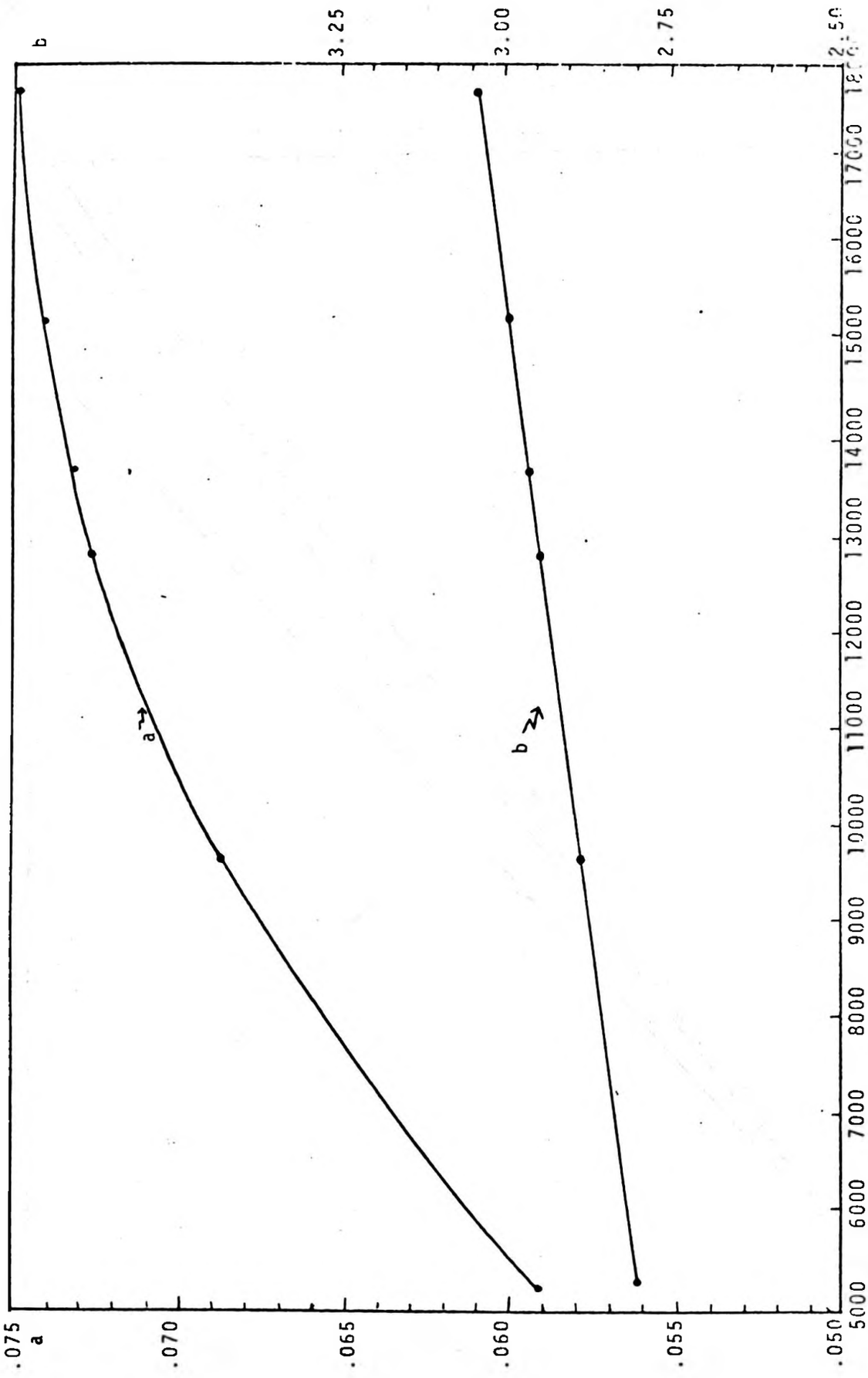


FIGURE 3 : VARIATION DE LA PENTE (a) ET DE LA CONSTANTE (b)

TABLE 1

NO	QUESTION	NRB	NRD	COUT($)
1	A1 01008, 01015; A2 WAVE($); A3 EXTENSIONAL & A1 & A2;	5200	2	$2.93
2	A1 DAMP($); A2 HYSTERE($), MATERIAL, INTERNAL, LAYER, COMPOSITE; A3 A1 & A2; --- INTERNAL & EXTERNAL & DAMPING;	5200	22	$4.86
3	01009 & 05014; --- A1 01008 & 01010; A2 GEAR($); A3 A1 & A2;	5200	65	$8.64
4	A1 FRICTION($); A2 HYSTERE($), MATERIAL, INTERNAL, LAYER, COMPOSITE; A3 A1 & A2; --- A1 01014; A2 03014, 04018, 06015; A3 A1 & A2;	5200	85	$8.18

NO	QUESTION	NRB	NRD	COUT($)
5	A1 01006; A2 02004, 02008, 02007; A3 03018; A4 MOVING($); A5 LOAD($), FORCE($); A6 A4 & A5; A7 A6 & A3; A8 05003, PROBAB($), STATIST($), 05017, 05029, 05005, 05038; A9 A1 & A2 & A7 & A8;	5200	520	$34.58
	A1 CABLE($), PENDULUM($), BOX($), CONTAINER($); A2 AERODYNAMIC($), FLUTTER($); A3 A1 & A2;			
	A1 01005, 01000, 01013; A2 01008 & 01004; A3 A1 & A2; A4 02003, 02004, 02005, 02001, 02007; A5 03012, 03018; A6 A3 & A4 & A5			
	A1 GEAR($); A2 ERROR($); A3 TRANSMISSION($); A4 DYNAMIC($), VIBRAT($); A5 A1 & A2 & A3 & A4;			
	A1 01000, 01003; A2 03018 & 04019; A3 A1 & A2;			

NO	QUESTION	NRB	NRD	COUT($)
6	A1 DAMP($); A2 HYSTERE($), MATERIAL, INTERNAL, LAYER COMPOSITE($); A3 A1 & A2; 01014, 03014, 04018, 06015;	5200	982	$60.00
7	03014 & 01014;	9600	29	$ 4.67
8	A1 MOVING($), TRAVEL($); A2 MATRI($); A3 TRANSFER($); A4 A2 & A3; A5 A1 & A4; 01006 & 03006 & 03013 & 04019	9600	90	$ 9.75
9	RANGE 5000, 12000; A1 MOVING($), TRAVEL($); A2 MATRI($); A3 TRANSFER($); A4 A2 & A3; A5 A1, A4;	12000	99	$ 9.77
10	A1 ROTAT($), 03006; A2 01006; A3 A1 & A2 A1 FRACT ($); A2 CRACK($); A3 A1 & A2;	1200	413	$32.30

NO	QUESTION	NRB	NRD	COUT($)
11	A1 SHAFT($), ROTOR($), BLADE($); A2 03006, 03007; A3 A1 & A2	12000	666	$51.23
	TURBO($);			
	MACHINE($);			
	COMPRESSOR($);			
	TURBINE($);			
	GEAR($);			
	BEARING($);			
	SEAL($);			
12	A1 01006; A2 02007, 02008; A3 03006, 03013, 03018; A4 A1 & A2 & A3; A5 ROTOR($); A6 A4, A5;	12000	1213	$88.83
	A1 01006 & 01010 & 03007; A2 BLADE($); A3 A1 & A2;			
	A1 BEARING($); A2 01009 & 01010; A3 A1, A2 A4 03013, 03015; A5 A3 & A4;			
	CASCADE($);			

NO	QUESTION	NRB	NRD	COUT($)
13	A1 MOVING($), TRAVEL($); A2 ELASTIC($); A3 FOUNDAT($); A4 A2 & A3; A5 03017, A4; A6 A1 & A5; A7 A6, LOCOMOTIVE($), TRACK($), RAIL($), TRAIN($);	12800	73	$ 8.69
14	A1 01017; A2 0622; A3 A1, A2;	12800	380	$29.60
	A1 VIBRAT($); A2 AXIAL($), SINUSOIDAL($), RANDOM($); A3 06005, 06007, 06009, 06020, 06019; A4 A1 & A2 & A3			
15	A1 BEAM($), MEMBRANE($), PLATE($), SHELL($); A2 IMPACT($), TRANSIENT($), RESPONSE($), FORCE($); A3 LIGHT, LASER, ELECTRON, HYPERVELOCTY, METEOR, METEORITE; A4 VISCOPLASTIC, PLASTIC, EARTHQUAKE, RANDOM, ACOUSTIC, THERMAL; A5 A1 & A2 & A3(NOT) & A4(NOT);	13600	350	$28.07

NO	QUESTION	NRB	NRD	COUT($)
16	A1 05013; A2 FINITE & ELEMENT; A3 A1, A2;	13600	614	$45.89
17	A1 BEAM($), RAIL($), TRACK($); A2 04003, 04005, 04007, 04008, 04009, 04010, 04014, 04016, 04017, 04019, 04020; A3 BRIDGE, BRIDGES; A4 A1 & A2 & A3(NOT);	13600	625	$50.66
	A1 BEAM($), 01006; A2 03017, ELASTIC($); A3 03018, FORC($), MOVING, TRAVEL($); A4 03005, BUCKL($), STABIL($); A5 A2 & A3; A6 A5, A4; A7 A1 & A6			
	A1 RAIL($), TRACK($), TRAIN($); A2 DYNAMIC($), VIBRAT($), RESPONSE($); A3 A1 & A2; A4 BRIDGE, BRIDGES; A5 A3 & A4(NOT);			
18	A1 GEAR($); A2 05013; A3 A1 & A2;	15200	4	$ 4.36
	A1 GENERAT($); A2 05013; A3 A1 & A2			
	A1 THREE($); A2 DIMENSION($); A3 05013; A4 A1 & A2 & A3;			

NO	QUESTION	NRB	NRD	COUT($)
19	RANGE 12000, 15200; A1 MOVING($), TRAVEL($); A2 MATRI($); A3 TRANSFER($); A4 A2 & A3; A5 A1, A4;	15200	45	$6.30
20	A1 BILINEAR; A2 MULTI & LINEAR; A3 PIECEWISE; A4 A1, A2, A3; --- A1 01003; A2 NONLINEAR; A3 A1 & A2;	15200	160	$14.31
21	A1 GEAR($); --- B1 PARAMETRIC($);	15200	177	$16.11
22	RANGE 12700, 17600; A1 LOCOMOTIVE($), TRACK($), RAIL($), TRAIN($); --- RANGE 15100, 17600; B1 GEAR($);	17600	88	$9.66
23	A1 TURBOPUMP($), PUMP($); --- B1 BUILDING($);	17600	194	$17.33
24	A1 BEAM($), PLATE($), SHELL($), COLUMN($); A2 05013, 05034, 06007; A3 A1 & A2;	17600	496	$39.45

NO	QUESTION	NRB	NRD	COUT($)
25	A1 05013; A2 STRUCT($), BEAM($), PLATE($), BRIDGE($); A3 A1 & A2;	17600	723	$57.37
	RANGE 15000, 17600; A1 TRANSFER($); A2 MATRI($); A3 A1 & A2;			
	A1 PROBABIL($), RANDOM($), STATISTIC($); A2 STRUCT($), BEAM($), PLATE($), BRIDGE($); A3 A1 & A2;			
	RANGE 15000, 17600; A1 MOVING($);			
	A1 TRUSS($), FRAME($);			
	A1 EARTH QUAKE($);			

TABLE 2

COMPARAISON ENTRE LE "COUT REEL"
ET LE COUT ESTIME PAR LA FORMULE

NO	NRB	NRD	COUT $ REEL	COUT $ EST.	ECART	%D'ERREUR
1	5200	2	2.93	2.92	-0.01	0.3
2	5200	22	4.86	4.10	-0.76	1.5
3	5200	65	8.64	6.64	-2.00	23 *
4	5200	85	8.18	7.82	-0.36	4.4
5	5200	520	34.58	33.48	-1.10	3.1
6	5200	982	60.00	60.74	+0.74	1.2
7	9600	29	4.67	4.88	+0.21	4.4
8	9600	90	9.75	9.06	-0.69	7
9	12000	99	9.77	10.02	+0.25	2.5
10	12000	413	32.30	32.51	+0.21	0.6
11	12000	666	51.23	50.63	-0.60	1.1
12	12000	1213	88.83	89.79	+0.96	1
13	12800	73	8.69	8.23	-0.45	5
14	12800	380	29.60	30.42	+0.82	2.7
15	13600	350	28.07	28.52	+0.45	1.6
16	13600	614	45.89	47.79	+1.90	4.1
17	13600	625	50.06	48.59	-1.47	2.9
18	15200	4	4.36	3.30	-1.06	24.3*
19	15200	45	6.30	6.33	+0.03	0.4
20	15200	160	14.31	14.84	+0.53	3.7
21	15200	177	16.11	16.10	-0.01	0
22	17600	88	9.66	9.65	-0.01	0.1
23	17600	194	17.33	17.60	+0.27	1.5
24	17600	496	39.45	40.25	+0.80	2
25	17600	723	57.37	57.28	-0.09	0.1
			642.94		15.78	2.4

TABLE 3

COUTS ESTIMES

NRB \ NRD	50	100	150	200	250	300	350	400	450	500
18000	6.81	10.57	14.32	18.07	21.83	25.58	29.34	33.09	36.85	40.60
19000	6.84	10.60	14.37	18.13	21.90	25.66	29.43	33.19	36.96	40.72
20000	6.87	10.64	14.41	18.18	21.95	25.72	29.49	33.26	37.03	40.80
21000	6.89	10.66	14.43	18.20	21.98	25.75	29.52	33.29	37.06	40.83
22000	6.91	10.68	14.45	18.22	21.99	25.76	29.53	33.30	37.07	40.84