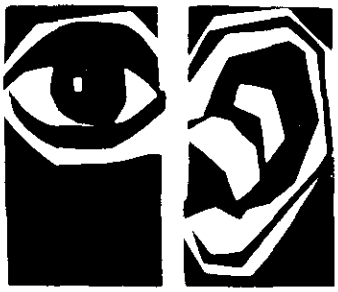
16 / A44 NCI Program
Scanning and Compression Project
Working Paper No. 4
September 18, 1973

EVALUATION OF IMAGE COMPRESSION FOR INSURANCE INDUSTRY DOCUMENTS



F. B. Wood
SDD Los Gatos
622-5265

R. B. Arps
SDD Los Gatos
622-5687

Confidential

Until December 31, 1978
For IBM Internal Use thereafter

IBM CONFIDENTIAL until December 31, 1978 Working Paper No. 4
thereafter IBM INTERNAL USE ONLY

Evaluation of Image Compression for Insurance Industry Documents

ABSTRACT

A "representative set" and a "additional set" of insurance industry documents were scanned and digitized in Rochester with a nominal 120 x 120 pel/inch resolution with 64 levels of grav scale. Their digital tape records were thresholded to black/white in Los Gatos.

The documents were then run through a series of black pel counting, prediction, run-length counting, and analysis programs to obtain statistics on:

(1) Compressed page distributions and document percent black vs. compressed record size (for market planners)

(2) Compressed scan line distributions and the ratio between low and high resolution storage requirements (for system architects)

(3) The impact on compression performance of changing dual base counters (for image product developers).

F. B. Wood
P. B. Arps

SUMMARY

The basic NCI algorithm used in the NCI Prototype compresses representative 180 kilo-byte documents to a range of 9 to 60 kilo-bytes with an average of around 30 kilo-bytes. A "pathologic" document that doesn't compress was found to require 181 Kilo-bytes of compressed storage. It has been examined and was found to be a copy on heat sensitive copier paper that had darkened with age.

The percent blackness for all the documents were plotted against their respective compressed record size. The points clustered into groups of documents of the same class, such as "applications", "claim reports", "change memos", etc. Further the clusters fell close to a slope line of 5 kilo-bytes per percent black. This rule of thumb of 5 kilo-bytes for each percent of blackness appears to hold for documents up to 15% blackness.

The compressed scan line analysis including both representative and additional documents shows that 1056-pel scan lines average to 176-bits per compressed line. 27.7% compress to 4-bits, 70.9% compress to values in the range 8 to 1056 bits, and 1.4% expand to the range 1060 to 2004 bits.

A subset of these documents was scanned at both 120 X 120 pel/inch and 240 x 240 pel/inch resolutions and analyzed. For equivalent threshold conditions, a document scanned at 240 X 240 pel/inch requires 2.6 times the compressed storage needed when it is scanned at 120 X 120 pel/inch.

Testing different dual-base counters with the 3-pel NCI predictor shows that programming different run length counters for separate classes of documents can give 7% more compression over using a fixed (10,6) counter.

TABLE OF CONTENTS

List of Figures

I. DESCRIPTION OF DOCUMENTS

A. Classification

Sixty-eight documents were scanned at nominal[1] 120 by 120 pels/inch resolution. The documents were numbered to indicate the class of document as shown below:

1. Applications
2. Inspection Reports
3. Motor Vehicle Reports (MVR's)
4. Renewal Notices which contain some form of correspondence from the policyholder
5. Policyholder correspondence
6. Change Requests
7. Correspondence from Agents
8. Automobile Claim Reports
9. Copies of Claims Drafts paid by Field Claims representatives
10. Polaroid pictures of damaged cars

The complete list of documents and copies of a few samples are included in Working Paper No. 3[2] . Three of the 21 representative documents are reproduced here, and the "pathologic" one from the set of 47 additional documents is reproduced here.

B. Examples of "Repesentatve Set" Documents

Figure 1:

The "Application for State Farm Automobile Insurance" represents the greatest detail and lowest compression ratio for a full page in the set of representative documents.

"Application For..." #1.0A (48)
Compression Ratio = 3.02 to 1.
Compressed from 1,440,384 to 476,972 bits
(59,622 bytes)

Figure 2:

The "Change of Address" form, when compressed individually comes close to having the average compression ratio as determined collectively for the

NCI Algorithm in the representative set of documents (8-1/2" x 11" field)

"Change of Address" #7.1A(61)
Compression Ratio = 6.90 to 1.
Compressed from 1,440,384 to 208,648 bits (26,081 bytes)

Figure 3:

The computer printout (5" x 3-1/2" data in an 8-1/2" x 11" field) of drivers license file data represents the document with the largest compression ratio within the representative set.

"A303101 13 .....M....." #3.0A (52)
Compression Ratio = 20.42 to 1.
Compressed from 1,440,384 bits to 70,541 bits (8,818 bytes)

C. "Pathologic Case" Example:

Figure 4:

The "Auto Underwriting Memo" which was scanned from a heat sensitive thermofax copy had sufficient dark background that it can be considered a pathologic case. This document had 21.5% black, while other "application" forms had between 11% and 15% black. The background of the faded thermofax paper added sufficient noise so that the document failed to compress. It took 1% more bits than the original to store the "compressed" output.

"Auto Underwriting Memo" #7.7B (45)
Compression Ratio - 0.99 to 1.
The compression algorithm expanded this document from 1,440,384 bits to 1,452,300 bits.

D. Typical Compressed Size for Different Insurance Forms

The "representative set" of documents, together with their insurance form numbers, are listed in Table I. The compressed sizes obtained with the NCI algorithm are included in the last column. The compressed size ranges between 8.9 kilo-bytes and 60.0 kilo-bytes for this set.

AG 4069 IL.2

APPLICATION FOR STATE FARM AUTOMOBILE INSURANCE

OFFICE COPY

Office Use

8879 691 13

VEH.	NEW	REINS.	TRANS	ADDED	QUALIFYING POLICY NO.	CLASS	REPLACES POLICY NO.
1			X				2743 895-E21-46 D.
2							DEC 11 1972

APPLICANT

NAME PLEASE PRINT	LAST NAME	FIRST NAME	MIDDLE NAME OR INITIAL
	RUIZ	TARZAN	F.

MAILING ADDRESS	NUMBER AND STREET	CITY	STATE	ZIP CODE
	PO Box 128	Hull	Ill	62343.

RESIDENCE If other than Mailing Address	NUMBER AND STREET	CITY	STATE	ZIP CODE	HAS APPLICANT LIVED IN THIS STATE 1 YEAR? YES	NO
	—				☐	☒

EXACT LOCATION (if residence, if no street number used): 8 mi East of Hannibal, Mo on Rt. 36. — COUNTY: Pike — TOWNSHIP: Salt Creek

FORMER ADDRESS (if in community less than 6 months): 8119 Dwen St. - Norfolk Va. — LENGTH OF RESIDENCY:

DURING THE PAST 6 YEARS, HAS THE APPLICANT OR ANY HOUSEHOLD MEMBER HAD AUTO INSURANCE CANCELLED, BEEN REFUSED ISSUANCE OR RENEWAL, OR RECEIVED NOTICE OF SUCH INTENT?	DURING THE PAST 5 YEARS, HAS THE APPLICANT, ANY HOUSEHOLD MEMBER, OR ANY REGULAR DRIVER: a. Had license to drive or registration suspended revoked or refused?	b. Been the driver in any automobile accident or loss?	c. Been convicted for forfeited bail for traffic violations?	DOES THE APPLICANT OR ANY REGULAR DRIVER HAVE ANY 1. PHYSICAL LIMITATIONS OR 2. MENTAL DEFECTS? IF YES, EXPLAIN IN REMARKS.
Yes ☐ No ☒	Yes ☐ No ☒	Yes ☒ No ☐	Yes ☐ No ☒	Yes ☐ No ☒

IF YES, WAS IT WITHIN THE PAST YEAR? Yes ☐ No ☐

MOST RECENT LIABILITY INSURER (COMPANY-EXPLAIN IF NONE)	POLICY NO.	FROM (MONTH-DAY-YEAR)	TO (MONTH-DAY-YEAR)
SF.			

VEH. 1

YEAR	MAKE AND MODEL	BODY TYPE	CYLS.	VEHICLE IDENTIFICATION NO.
72	FORD - GALAXIE	4dr. sed.	8	2N54H126471

AIR CONDITIONING YES / NO	MO. AND YR. PURCHASED	NEW / USED	COST INCL. EQUIP.	AMT. OWED	OTHER INSURANCE	FOR HIGH PERFORMANCE VEH. ONLY: ENGINE DISPLACEMENT	C.C. OR C.I.D.	HORSE-POWER	UTILITY VEH. COST NEW
(YES)	9 72	X	$ 4960	$ 3000	—	—		—	$ —

LIENHOLDER: Bank of Virginia — MAILING ADDRESS: 21st & Granby St. - Norfolk Va. — ZIP CODE: 23517.

VEH. 2

YEAR	MAKE AND MODEL	BODY TYPE	CYLS.	VEHICLE IDENTIFICATION NO.

AIR CONDITIONING YES / NO	MO. AND YR. PURCHASED	NEW / USED	COST INCL. EQUIP.	AMT. OWED	OTHER INSURANCE	FOR HIGH PERFORMANCE VEH. ONLY: ENGINE DISPLACEMENT	C.C. OR C.I.D.	HORSE-POWER	UTILITY VEH. COST NEW
			$	$					$

LIENHOLDER — MAILING ADDRESS — ZIP CODE

RATING

NO. CARS IN HOUSEHOLD	NO. INS. BY STATE FARM	EXISTING DAMAGE OR MODIFIED VEH. 1	NO	YES-DESCRIBE	VEH. 2	NO	YES-DESCRIBE
1	1		X	—			

VEH.	TERR. NO.	IN CITY	BUSINESS USE	FARM RATE	DRIVEN TO AND FROM WORK: MILES ONE WAY	DRIVEN TO AND FROM WORK: AVERAGE WEEKLY	ESTIMATED ANNUAL MILEAGE	2 CAR RULE	DRIVER TRAINING	GOOD STUDENT	CHILDREN AT SCHOOL	RATING GROUP	CLASS	ACC. GRP.
1	40	No	—	—	0	0	8000 - 10,000	—	—	—	—	7	180	-
2														

COVERAGES

The insurance applied for is only for the coverages indicated by specific premium entry. If premium cannot be entered, check boxes to indicate coverage requested.

	VEH. 1 AMEND LIMITS TO	VEH. 2	VEHICLE 1 MEMB.	VEHICLE 1 PREMIUM	VEHICLE 2 MEMB.	VEHICLE 2 PREMIUM
BIPD 25/50/15	50/100/25		AB ☒ T	34 20	AB ☐ XX	
MED. PAY. LIMITS	5000		C ☒ XX	4 90	C ☐ XX	
COMPREHENSIVE DEDUCTIBLE AMOUNTS	0		D ☒ T	12 80	D ☐	
80% COLLISION			F ☐		F ☐	
COLLISION DEDUCTIBLE AMOUNTS	50		G ☒ T	32 30	G ☐	
EMERGENCY ROAD SERVICE			H ☒ XX	1 30	H ☐ XX	
RENTAL REIMBURSEMENT			R ☐		R ☐	
UNINSURED MOTOR VEHICLE	☐ 10/20 Limits ☒ Same as BI	☐ 10/20 Limits ☐ Same as BI	U ☒ XX	5 40	U ☐ XX	
AUTO DEATH INDEMNITY SPECIFIC DISABILITY			S ☐			
LOSS OF EARNINGS			Z ☐ XX		COMPLETE S & Z SECTION	
ENDORSEMENTS			TOTALS	90.90		

S & Z SECTION

FULL NAME OF PERSONS TO BE INSURED (S-Z) — ONLY PERSONS IN APPLICANT'S HOUSEHOLD ARE ELIGIBLE	S-PRINCIPAL SUM AMOUNT	PREM.	Z PREMIUM

MPP ACCOUNT NO.	AGENT'S CODE STAMP
DATE AND TIME OF APPLICATION: MO. 12 DAY 7 YR. 72; 3 ☐ AM ☒ PM	M. BREDER 8728 HUNSAKER 27
AGENT'S CHECK $ —	
APPLICANT'S CHECK $	REMITTANCE TO BE APPLIED TO: VEH. 1 $ — VEH. 2 $
TOTAL BALANCE DUE $ 0	

REMARKS

SIGNATURES

BINDER — EFFECTIVE DATE: VEH. 1 12-7-72; VEH. 2

☒ STATE FARM MUTUAL AUTOMOBILE INSURANCE COMPANY ☐ STATE FARM FIRE AND CASUALTY COMPANY

of Bloomington, Illinois, hereby binds, as of the requested effective date, the insurance applied for, subject to all of the terms and conditions of the automobile policy in current use by such company for a period of 30 days from such effective date. Issuance of the policy applied for voids this binder.

AGENT'S SIGNATURE X Milton E Breder

I hereby apply for the insurance indicated and represent (1) I have read this application, (2) the statements hereon are correct and (3) I am the sole owner of the described automobile(s) except as otherwise stated.

IT IS UNDERSTOOD AND AGREED THAT NO INSURANCE IS EFFECTIVE HEREUNDER (A) UNLESS THE BINDER IS COMPLETED DESIGNATING THE COMPANY ACCEPTING THIS APPLICATION AND SIGNED BY AN AUTHORIZED AGENT OF SUCH COMPANY OR (B) UNTIL THE DATE OF THE POLICY OR BINDER ISSUED BY THE COMPANY ACCEPTING THIS APPLICATION.

"If a policy is issued to me by State Farm Mutual Automobile Insurance Company, effective as of the date such policy is issued I hereby appoint the President of such Company, or such person as he may designate, to vote my proxy at any and all meetings of members at which I am not present in person or by subsequent proxy, this proxy to remain in force until expressly revoked."

APPLICANT'S SIGNATURE X Tarzan F. Ruiz

Fig. 1. "Application for State Farm Insurance", Doc. #1.OA

CHANGE OF ADDRESS

IMPORTANT: If any policy has a different named insured, indicate the name immediately after the policy number.

NAME OF INSURED: George Thomas Current

NEW ADDRESS: ☒ PERMANENT ☐ TEMPORARY, HOW LONG? _____

RESIDENCE 215 E. Bair St Olney Ill
(NUMBER STREET CITY STATE ZIP CODE COUNTY, TOWNSHIP)

MAILING _____
(NUMBER STREET CITY STATE ZIP CODE)

EFFECTIVE DATE: 12-6-72 MPP ACCOUNT NO: _____

16 DEC 08 1972 POSTDATE (AGENT)

☒ **AUTO**

POLICY NUMBERS	CLASS CHANGES FROM	CLASS CHANGES TO	REASON	FARM RATED?	TERR. NO.
6585-903-F09BD				YES NO	
6651 543-A14-13C				YES NO	INSIDE CITY LIMITS?
				YES NO	YES NO
				YES NO	

☐ **LIFE** POLICY NUMBERS

_____ _____

_____ _____

_____ _____

_____ _____

☐ **FIRE** INSURED MOVED: ☐ INTO ☐ WITHIN ☐ OUT OF MY TERRITORY

POLICY NUMBERS	CANCEL EFFECTIVE	REWRITE POLICY (ATTACH APP)	ENDORSE REMOVAL (COMPLETE)
		☐	☐
		☐	☐
		☐	☐

CONSTRUCTION	YEAR BUILT	TYPE OF ROOF	OCCUPIED AS	NUMBER OF FAMILIES	BOARDERS OR ROOMERS
PROTECTION CLASS	DISTANCE TO HYDRANT FEET	TO FIRE DEPT. MILES	IF OUTSIDE CITY, DISTANCE FROM CITY LIMITS MILES	ANNUAL CONTENTS RATES AT NEW LOCATION FIRE	EC
NAME OF SERVICING FIRE DEPARTMENT		NAME OF FIRE DISTRICT (WHERE APPLICABLE)			

NUMBER SQ. FEET	FIRST FLOOR	SECOND FLOOR	OCCUPIED BY	OWNER	TENANT	VACANT	IF VACANT, EXPLAIN

(1) FOR NON-RESIDENTIAL CONTENTS REMOVAL, USE APPLICATION F7-865 (2) ATTACH SNAPSHOT (AS REQUIRED BY REGIONAL OFFICE)

☐ **HEALTH** POLICY NUMBERS

_____ _____

_____

DATE 12-6-72 AGENT'S NAME AND CODE R C McDowell 8456

MG 5941

Fig. 2. "Change of Address", Doc. #7.1A

L200-6521-2200 00620
A303161 13 M 72-12-05 12-07-72

THE FOLLOWING INFORMATION IS FURNISHED FROM THE DRIVERS LICENSE FILE OF THE PERSON IDENTIFIED ABOVE PURSUANT TO THE PROVISIONS OF THE ILLINOIS VEHICLE CODE.

CTIS L LUCY
800 MARKLAND L200-6521-2200
SALEM 62881 BIRTH DATE

SEX	FT.	WT.	HAIR	EYES	TYPE	ISS DATE	CLASS	RESCRICT	EXP DATE
M	68	205	BRWN	BRWN	1	08 09 71	A*	100	07 14 74

TYPE ACTION	DATE OF ARREST	DATE OF ACTION	DESC OF ACTION	ACC OR DOCKET NO.	TERM DATE OF SUSP	STOP IN EFFECT
99 YRDW	092468	093068	1070000	0001706	00000	00
99 FTC	041770	041770	1061000	0116994	00000	00

Fig. 3. "Drivers License File" (printout), Doc. #3.0A

(E)

AUTO UNDERWRITING MEMO

DATE 10-9-72 FROM: [illegible] TO: [illegible]

POLICY NUMBERS: 1549911130 NAME [illegible]

ADDRESS: [illegible]

PLEASE REPLY TO ITEMS CIRCLED

1. Contact _____ and secure details of accident(s)/violation(s) in which he/she was involved on _____.

2. The change of name recently requested will require a fully completed and signed application.

3. Obtain a release of credits from _____.

4. Forward a driver training/good student certificate for _____.

5. We understand _____ has a health problem. In order to properly review the risk we need to ask that a medical report be reviewed by our Medical Department. Please have him and/or her fill in the COMPLETE NAMES AND ADDRESSES of doctors who have treated him and/or her for _____ problem and/or she has been confined recently on the enclosed medical authorization form. If he and/or she has not been confined recently, then the last doctor he and/or she visited. After completion, have the forms signed and return them to us. (If individual is a minor have parents sign.)

6. Furnish driving directions to Insured's residence, plus references.

7. We understand the insured has moved. Obtain current address and zip code.

8. [illegible] of the insured during the past year and length of time at each.

9. The vehicle insured under this policy was a total loss on _____. What disposition is to be made of this policy?

10. Forward drivers license number, date issued and date of birth for _____.

11. The driver license number for _____ is incorrect. The State has no record for the number given. Forward correct one.

12. The insured reportedly is not a citizen of the U.S. Does he/she have a permanent visa for [illegible]? Will he/she be in the U.S. 12 months?

13. Driver information on [illegible]
- Full Name
- Date of Birth
- Driver License
- Occupation
- Prior Cancellation [illegible]
- Any physical defects?
- Dates and details of prior accidents and violations

14. Review indicates class _____ applies because _____

[illegible]

REPLY BELOW. 2nd REQUEST—PLEASE REPLY OCT 20 1972

[illegible] forms to drop off [illegible] information [illegible] Jayce will call insured again this week to get the information & will follow up in one week to make sure information obtained

DATE 12/4/72 SIGNATURE [illegible]

Fig. 4. "Auto Underwriters Memo", (faded copy), Doc. #7.7B

TABLE I

State Farm Insurance
Representative Set of Documents
Compressed by NCI Algorithm
(120 X 120 pel/inch Resolution)

Document Number	Insurance Form No.	Description	Compressed Size (Kilo Bytes)
1.0A*	AG 4069 IL.2	Application "Ruiz"	60.0
1.0B*	AG 4069 IL.2	Bottom of Application "Ruiz"	18.3
2.0A*	6033	Automobile Insurance Report "Weil"	55.0
2.0B*	6033-F	Backside of Auto Insurance Report "Weil"	28.8
3.0A		Computer Printout from MV License File	8.9
4.0A*	G 54864.1	Renewal Notice "Smith"	12.1
4.0B*	G 54864.1	Backside of Renewal Notice	11.4
5.0A*	G 54844.2	Follow-up Renewal Notice "Kroll"	13.8
5.0B		Handwritten letter "Kroll"	17.4
5.1A		Handwritten letter "Mathes"	34.5
5.1B		Second page of handwritten letter	19.5
6.0A	G 4178.2	Change Request "Vlahan"	44.6
7.0A*	AG 4709.12	Change Memo "Belz"	33.3
7.1A	MG 5941	Change of Address "Current"	26.2
7.2A		Policy holder Review "Burns"	20.3
7.3A	N 4371.4	Inter-Office Correspondence "Keeran"	13.5

Document Number	Insurance Form No.	Description	Compressed Size (Kilo Bytes)
7.4A	AG 4070	Binder for SF MVI "Dunes"	28.4
7.5A*	MG 4517.6a	Agents MEMO (change) "Pomack"	39.4
8.0A*	G 4684n	Automobile Claim Report "Clark"	55.7
9.0A*		Claim Draft "Mesbit"	18.0

---

* This document and others of the same type have been plotted on the sheet "Per Cent Black vs. Compressed Image Size" to show clustering. (Fig. 7)

II. SAMPLE DISTRIBUTION FOR NCI ALGORITHM

Each of the 21 documents from the "representative set" and each of the 47 documents in the "additional set" were run through the compression calculating programs for the NCI Algorithm. The results are tabulated in Table II and III. Those tables are for images obtained by scanning on the PIDAS scanner in IBM Pochester and computer thresholding at Los Gatos. The document numbers are the same as in Working Paper No. 3.

The second column in Table II is the compression ratio obtained with the NCI algorithm. The "bound" shown in parentheses is the theoretical limits of compression for that document with the NCI predictor. The fourth column "Efficiency" gives the efficiency of the (10,6) dual base run length codes when used with the predictor output for the given document. The fifth and sixth columns ("Compressor Kilo Bits/Kilo-bytes) give the compressed image size in kilo-bits with the size in kilo-bytes in parentheses.

The distribution of compressed documents sizes are plotted in Fig. 5.

Gaussian curves for the same mean and variance are also plotted for comparison.

The compressed image sizes are plotted in Figure 6 for the NCI algorithm for both the "representative set" and the "addition Set". A bar has been added to show the image size of the raw data. The "pathologic case" is also included.

TABLE II

STATE FARM DOCUMENTS
Representative Set of 21 Documents
1,440,384 kilo bits
(180 kilo-bytes)

S1D5U0 NCI ALGORITHM OK

Doc. Num.		Comp. (bound)	Eff.%	Compressed Kbits (Kbytes)
1.0A	48	3.0 (3.6)	85.0	477 (59.6)
1.0B	49	9.9 (11.2)	87.8	146 (18.3)
2.0A	50	3.3 (4.0)	82.7	440 (55.0)
2.0B	51	6.3 (7.4)	84.5	230 (28.8)
3.0A	52	20.4 (23.9)	85.6	71 (8.9)
4.0A	53	14.8 (17.5)	84.5	98 (12.2)
4.0B	54	15.8 (18.3)	87.8	91 (11.4)
5.0A	55	13.0 (15.3)	85.3	110 (13.8)
5.0B	56	10.3 (11.5)	90.1	139 (17.4)
5.1A	57	5.2 (5.8)	89.6	276 (34.5)
5.1B	58	9.3 (10.4)	89.2	156 (19.5)
6.0A	74	4.0 (5.7)	70.7	357 (44.6)
7.0A	60	5.4 (6.3)	85.3	266 (33.2)
7.1A	61	6.9 (7.9)	87.0	209 (26.2)
7.2A	62	8.8 (10.0)	88.8	162 (20.3)
7.3A	63	13.4 (15.0)	88.8	108 (13.5)
7.4A	64	6.4 (7.2)	88.4	227 (28.4)
7.5A	65	4.6 (5.4)	84.9	314 (39.2)
8.0A	66	3.2 (3.8)	85.5	445 (55.6)
9.0A	67	10.0 (11.6)	86.3	144 (18.0)
10.0A		13.6 (16.9)	80.4	107 (13.4)

Min. 3.0A 71 Kbits 8.88 Kbytes
Max. 1.0A 477 59.7

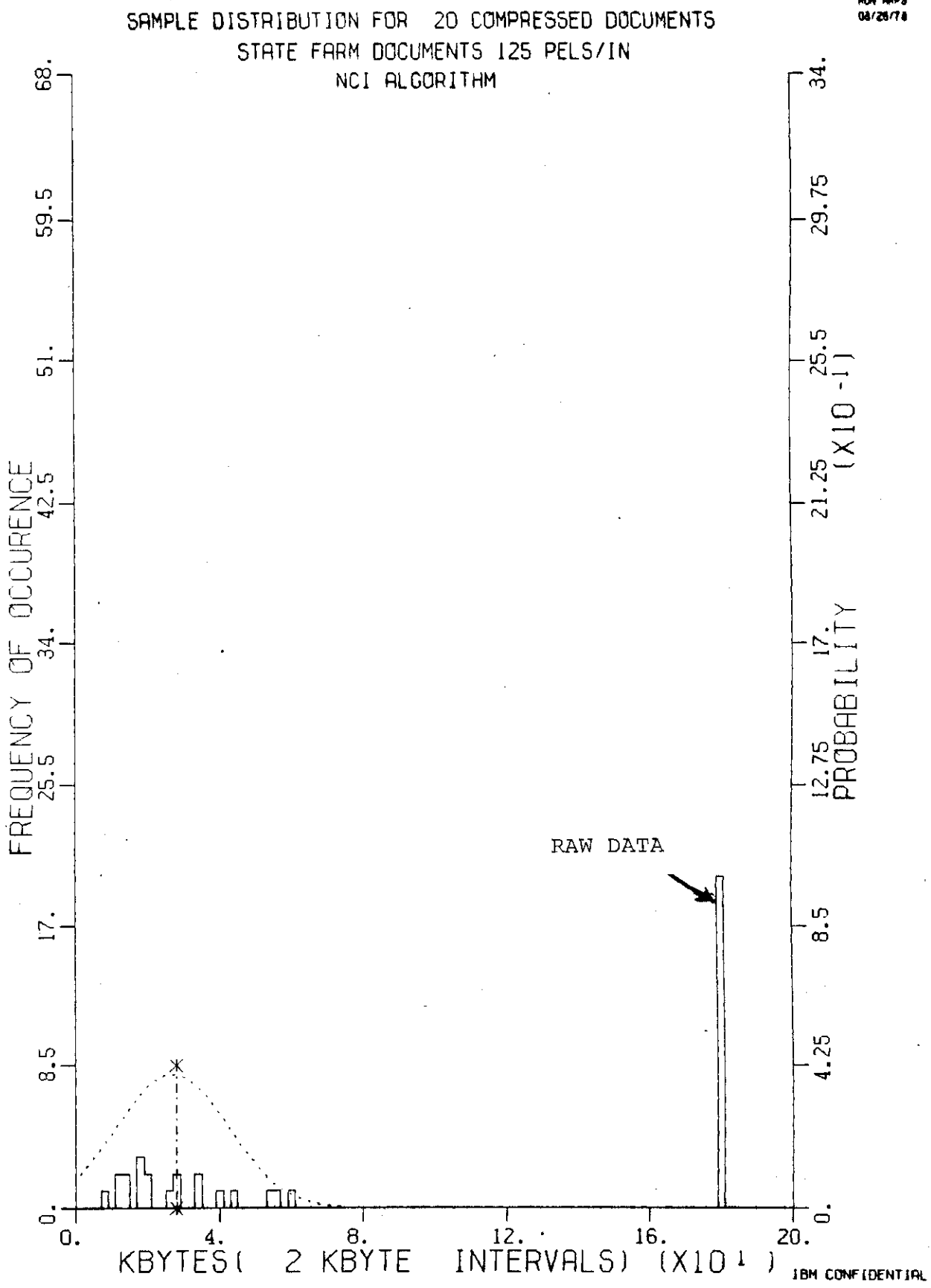


Fig. 5. Sample Distribution of Compressed Document size for Representative Set using NCI Algorithm (2 kilo-byte intervals)

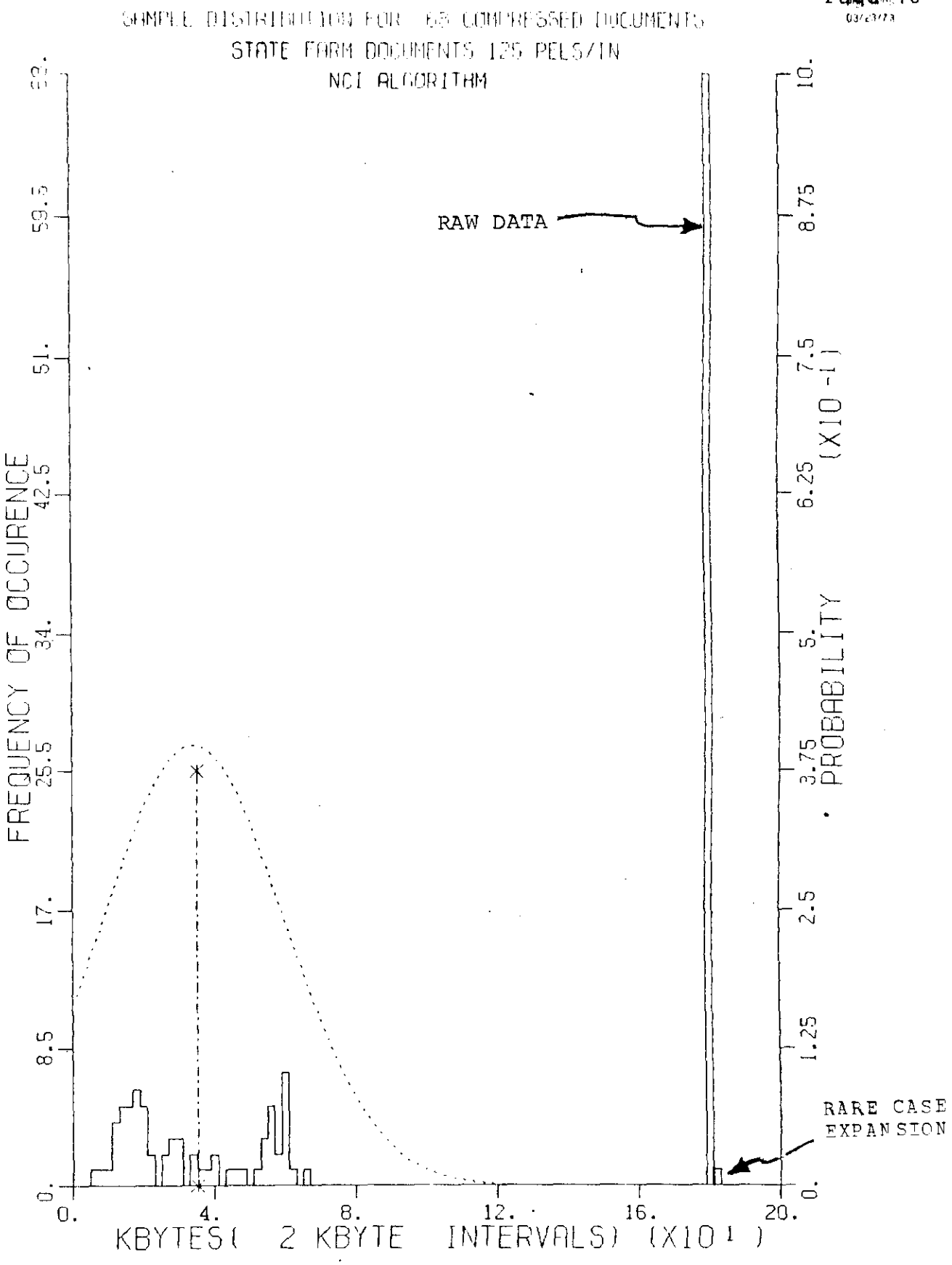


Fig. 6. Sample distribution of Compressed Document size for Representative plus Additional Set using NCI Algorithm. (2 Kilo-byte intervals)

TABLE III

STATE FARM DOCUMENTS
Additional Set of 47 Documents

Doc. Num.	NCI ALGORITHM Comp. (Bound)	Eff. %	Compressed K-Bits
1.1A	3.2 (3.8)	85.5	446
1.1B	11.6 (13.2)	87.9	124
1.2A	3.2 (3.8)	85.4	448
1.2B	13.3 (15.1)	87.7	109
1.3A	2.0 (3.5)	85.1	487
1.3B	10.6 (12.1)	87.7	136
1.4A	3.0 (3.5)	85.3	484
1.4B	11.0 (12.5)	88.0	131
1.5A	3.2 (3.7)	85.8	454
1.5B	10.5 (11.8)	88.5	137
1.6A	3.0 (3.5)	85.8	485
1.6B	14.6 (16.8)	87.0	96
1.7A	3.0 (3.5)	85.5	479
1.7B	9.2 (10.5)	87.5	157
1.8A	3.3 (3.9)	85.0	440
1.8B	27.1 (32.0)	84.6	53
1.9A	2.8 (3.2)	85.2	524
1.9B	9.7 (11.1)	87.6	149
2.1A	3.4 (4.1)	82.1	429
2.1B	6.8 (8.2)	83.7	211
2.2A	3.1 (3.8)	83.3	460
2.2B	5.1 (5.9)	86.0	282
2.3A	2.9 (3.5)	83.1	502
2.3B	4.7 (5.4)	86.0	309
6.1A	6.6 (7.8)	85.1	218
6.1B	9.3 (11.0)	84.9	154
6.1C	11.3 (13.1)	86.2	128
8.1A	3.3 (3.8)	86.0	438
8.2A	3.8 (4.6)	83.9	377
8.3A	3.1 (3.6)	84.9	473
8.4A	3.4 (4.0)	85.9	422
8.5A	3.0 (3.5)	86.1	483
8.6A	3.1 (3.6)	85.2	470
5.1A	18.0 (21.3)	84.5	80

		NCI ALGORITHM		
Doc. Num.		Comp. (Bound)	Eff. %	Compressed K-Bits
5.1B		10.0 (11.4)	87.6	145
5.2A		6.0 (6.5)	91.6	241
5.2B		8.2 (9.4)	87.8	175
5.2C		9.0 (10.2)	88.1	160
5.2D		8.3 (9.8)	84.3	174
9.1A		11.3 (13.2)	85.3	128
9.2A		12.7 (15.1)	84.5	113
7.6A		6.1 (7.0)	86.9	238
7.6B		16.0 (18.9)	84.9	90
7.7A		6.0 (7.1)	83.5	242
7.7B	**	0.99(1.40)	70.7	1,452
7.8A		4.6 (5.4)	86.0	313
7.8B		4.0 (4.6)	86.1	365

** Image that expanded

III. RELATION BETWEEN PERCENT BLACK IN DOCUMENT AND THE RESULTANT COMPRESSED IMAGE SIZE

The complete set of 68 documents were run through two computer programs, one program counts the number of blacks pels in the black/white source document. The other uses the NCI Algorithm predictor and dual base (10,6) run length codes to calculate the compressed image size.

The percent black was then plotted versus the compressed image size in Figure 7. Only the documents for which there were two or more of the same type are included in the figure. It was apparent that documents of similar type clustered about particular regions on the percent black vs. compressed image size plot. In figure 7 the clusters are labelled with the class of document and the insurance industry form numbers.

The dashed line represents a slope of 5 kilo-bytes of compressed image per 1% black in a 180 kilo-bytes source document.

IV. DISTRIBUTION OF COMPRESSED SCAN LINE SIZE FOR NCI ALGORITHM

All documents in the "representative set" and the "additional set", a total of 68 documents, were compressed using Frank Nemec's assembly language implementation of the NCI algorithm. The distribution of compressed scan line lengths for 68 X 1364 = 92752 lines is plotted in Figure 8. The frequency distribution for different compressed line lengths ran between 1 and 25732, which is too large a range to show in a linear scale. Therefore, three different scales are used in Figs. 8, 9, and 10. 27.7% of the scan lines compress to 4-bits and 70.9% compress to values in the range of 8 to 1056 bits. The remaining 1.4% expand to the range 1060 to 2004 bits.

The scale resolution has been increased in Fig. 9 to show more detail of the distribution and to compare the actual distribution with an empirically fitted geometric mass distribution function, P(J):

$$P(J) = 0.277 \qquad \text{for } J=1$$

$$= 0.011(0.986)^{(J-2)} \qquad \text{for } J=2,3.....,501$$

where J is the number of compressed half-bytes, and L(bits) = 4 (J) (half-bytes)

The scale of Fig. 9 is still too small to see the detail of the few lines that expanded to almost twice input size. The same data is plotted in Fig. 10 for a finer scale to show the distribution of the 1.4% of the lines that expanded.

V. COMPARISON OF LOW AND HIGH RESOLUTION DOCUMENTS

The "representative set" of documents were rapidly scanned by the Los Gatos Raster Scanner at 120 X 120 pel/inch and 240 X 240 pel/inch resolution under equivalent threshold conditions to quickly obtain a comparison of the same documents for low and high resolution.

A complete analysis of Table IV indicates that the source black pels for high resolution is slightly

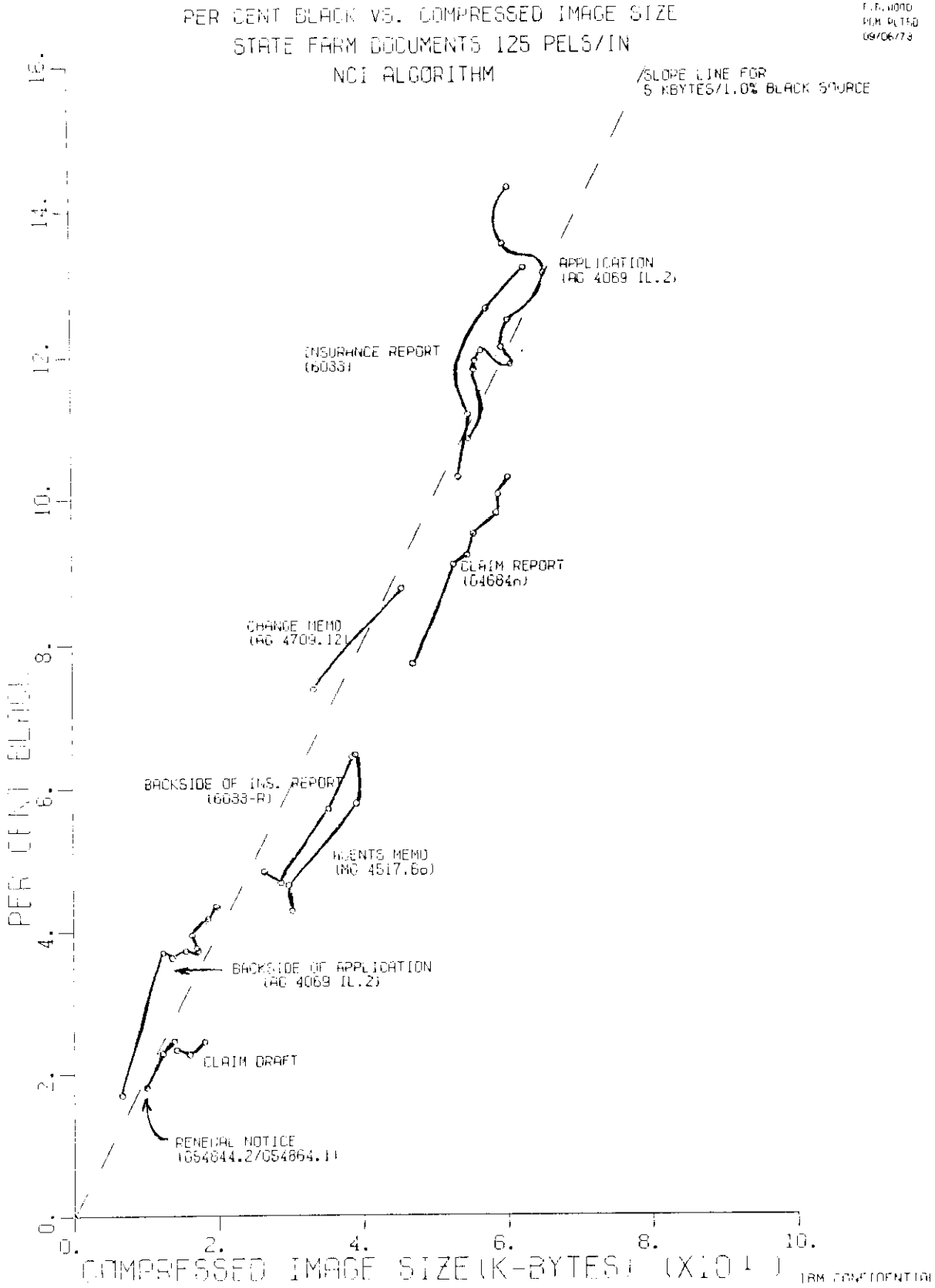


Fig. 7. "Percent Black vs. Compressed Image Size"

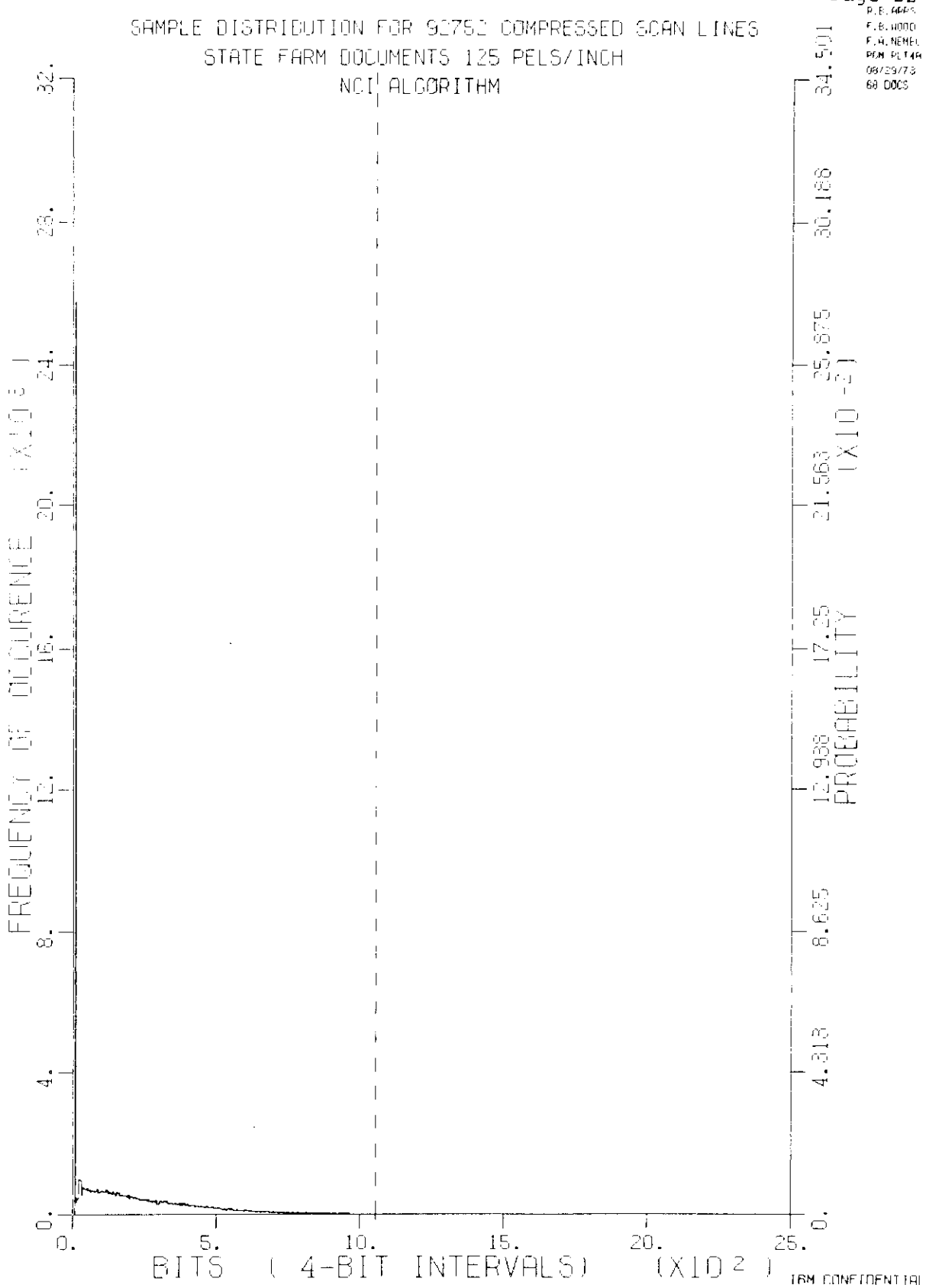


Fig. 8. Sample Distribution for 92752 Compressed Scan Lines (68 documents) Full Scale to Show Maximum

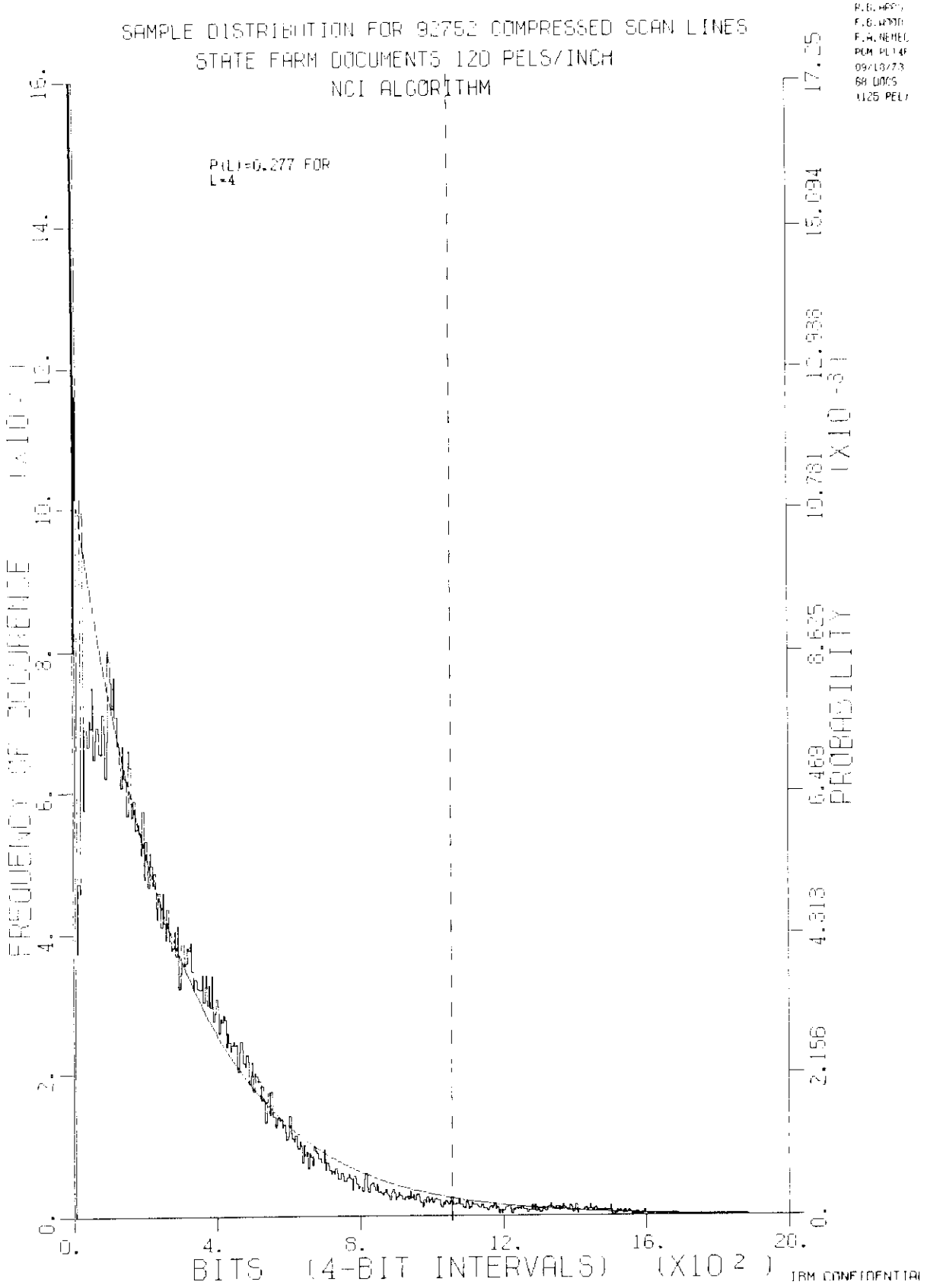


Fig. 9. Data for Fig. 8 at larger scale.
L=4 point is truncated.
Curve from geometric model is also included.

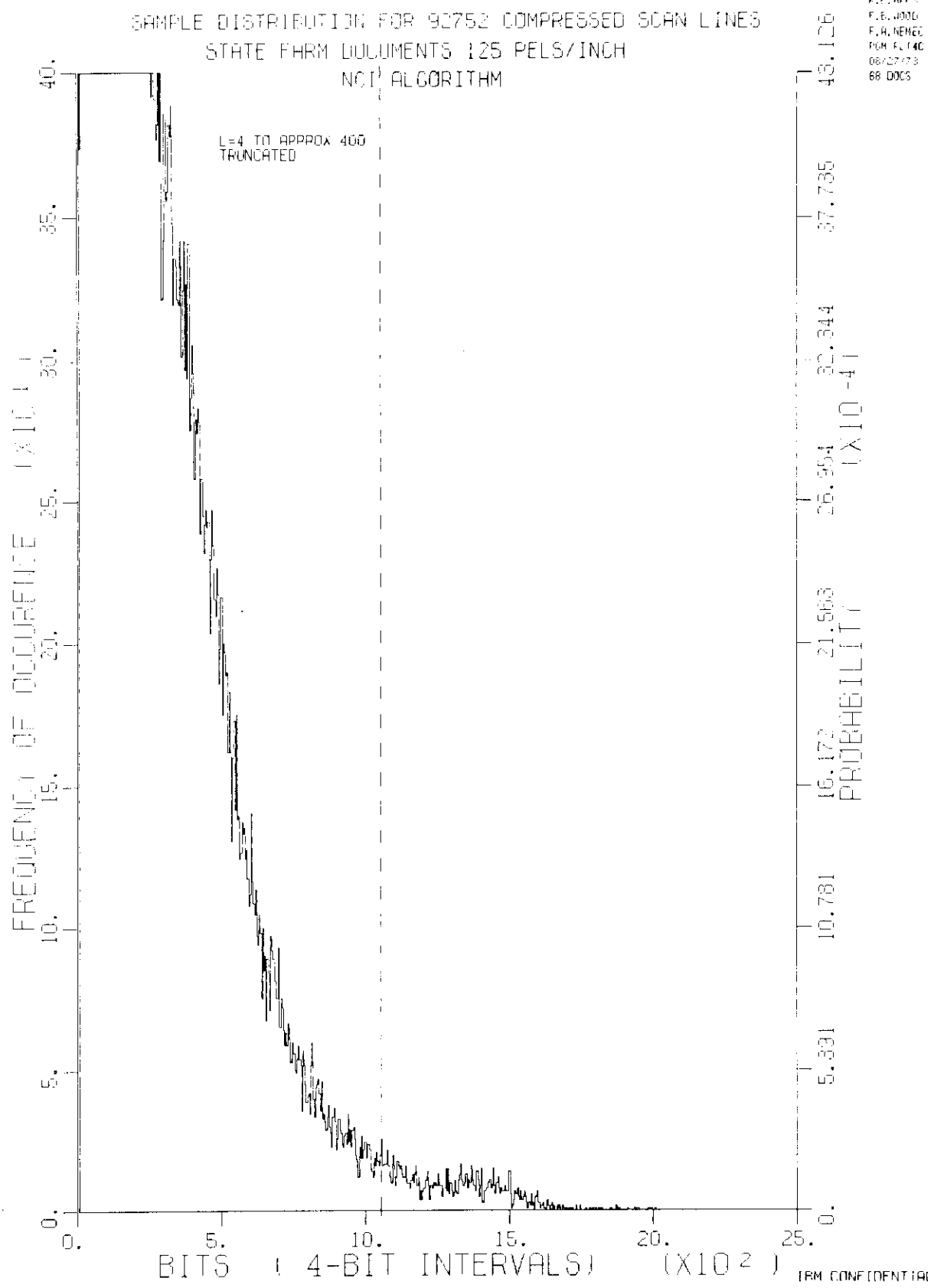


Fig. 10. Expanded scale to show the few lines that expanded.

higher than four times the black pels at low resolution. This is reasonably close to the theoretical factor of four that would be anticipated from doubling the resolution in both dimensions.

The compressed record sizes at high resolution are plotted against the compressed record sizes at low resolution in Fig. 11. A line indicating a 2.6X ratio between low and high resolution compressed size has been added as a rule of thumb. (Note that the theoretical lower bound is a ratio of 2.3X).

This data is slightly different than the PIDAS scanned, dynamic thresholded data used in the rest of this report. However, the ratio should be very indicative of the algorithm's resolution response.

VI. NCI OPTIMIZED ALGORITHM

The 21 documents in the "representative set" were concatenated as one large document of dimension 8 1/2" X 231" in order to explore optimal dual base counting schemes and obtain maximum coder efficiency. This concatenated document was run through the NCI 3-bit predictor followed by a program which computed the compressed document size for all possible dual base run length coders having 2-bit, 3-bit, 4-bit, or 5-bit counters.

The dual base counters follow the formula:

$$\text{Count} = a_0 \ (p^0 n^0) + a_1 \ (p^1 n^0) + a_2 \ (p^1 n^1) + a_3 \ (p^1 n^2) + \text{-------} + a_i \ (p^1 n^{i-1})$$

For 2-bit counters $p + n = 2^2 = 4$, giving $2^2 - 1 = 3$ possible dual base counter systems:

$(p,n) = (1,3), (2,2), (3,1)$

For 3-bit counter $p + n = n^3 = 8$; $2^3 - 1 = 7$ and the possible dual base counting schemes are:

$(p,n) = (1,7), (2,6), (3,5), (4,4) (5,3), (6,2), (7,1)$

The results are plotted in Figure 12. The compression bound for this concatenated data set for the NCI predictor is 7.693. The largest compression ratio was obtained for the (5,3) - code, giving a compressed image size of 4,251,882 bits with a compression ratio of 7.114, which is an efficiency of 92.5%.

A number of other codes had fairly good efficiency. They are listed in Table V.

data is padded out to an integral byte boundary, if necessary.

When using a variable word length, there must be a way of separating the individual code words out of a data stream. This is accomplished by restricting the low order digit (4 bits) of each code word to take on only the values 0 - 9. The high order digits (4 bits each) are restricted to the hex values A-F. The highest order digit is prevented from taking on the hex value A (it is restricted to B-F) since a code word of a hex A is given a different interpretation. Normally, the code represents the length of a run of white pels (non-errors in the error image). It is assumed that a black pel (error) follows the run of white. Two black pels in succession are represented by a white (non-error) run of zero length.

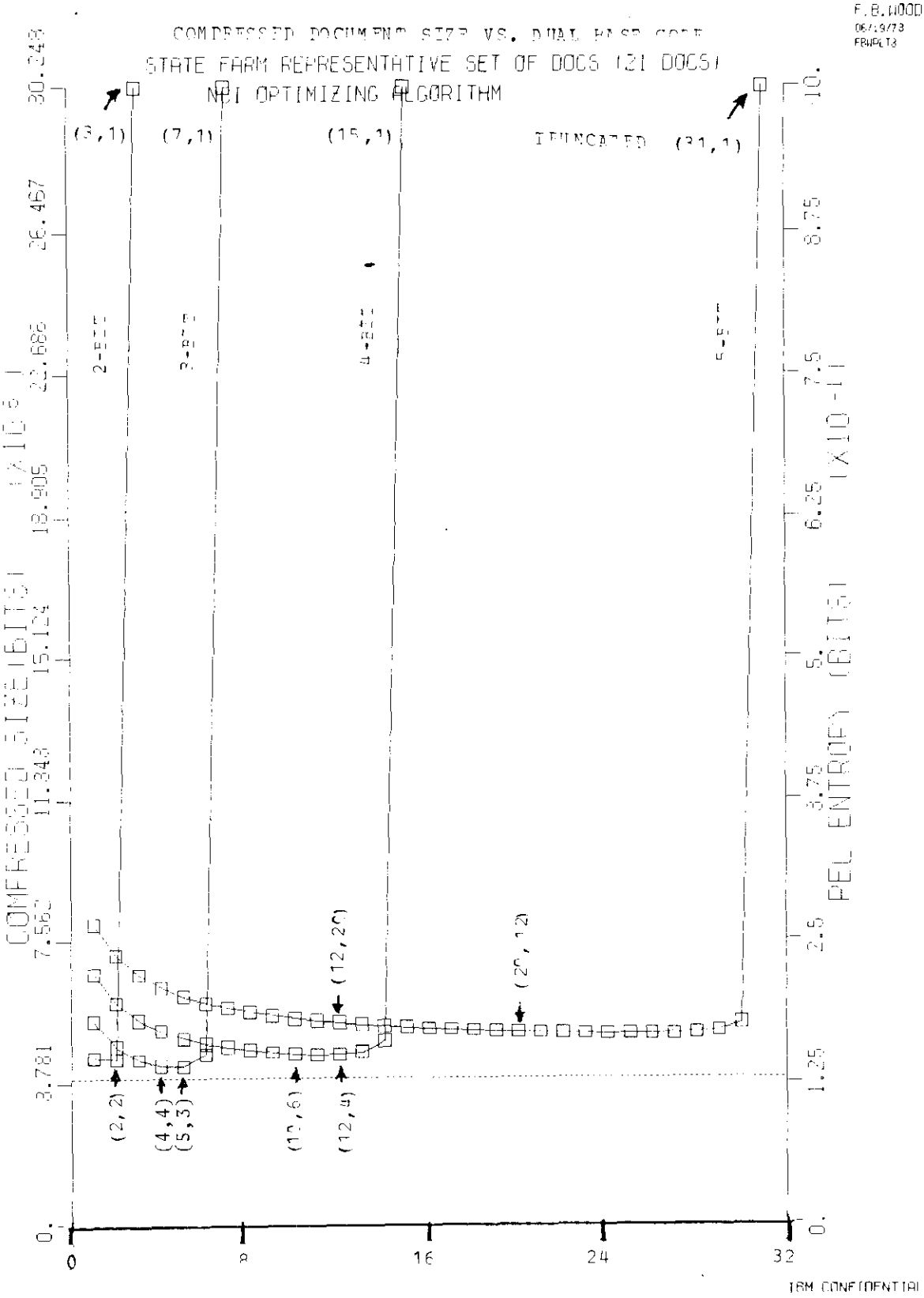
No ordinary code word is permitted to begin with a hex A. The code word for the last run in a scan line that does not have a black pel (error) at its end uses the hex A. This code of A indicates that the rest of the current scan line is all zeros (non-errors). Note that a blank line generally produces a line of non-errors and is indicated by a code of A. The total code, along with the run-lengths that can be encoded by the code words of various lengths are shown in Figure 13.

The compressor can be used with less than full pages (fields of images). A code word consisting of 4 hex F's is used to indicate the end of the partial image field. If padding is required to end on an integral byte boundary, five F's are used.

Decompression starts by separating out each code word. The encoded run length is used to construct a run of the proper length of all white followed by a black pel. This is repeated scan line by scan line to reconstruct the error image. The error image is then used to correct outputs from the predictor.

Starting with the all white scan line assumed to precede the first scan line, the first pel of the page is of course predicted to be white, which is compared with the first pel of the error image to reconstruct the first pel of the original image. This pel is then stored to be used for later predictions. The above

F.B.WOOD
06/19/73



P in (P,M) DUAL BASE CODE

Fig. 12. Compressed Document Size vs. Dual Base Code for Representative Set with NCI Predictor for all 2-bit, 3-bit, 4-bit, and 5-bit Dual Base Counters

process is repeated until the whole original image is reconstructed. Both the run length decoding and the prediction proceed concurrently.

APPENDIX A - NCI Algorithm

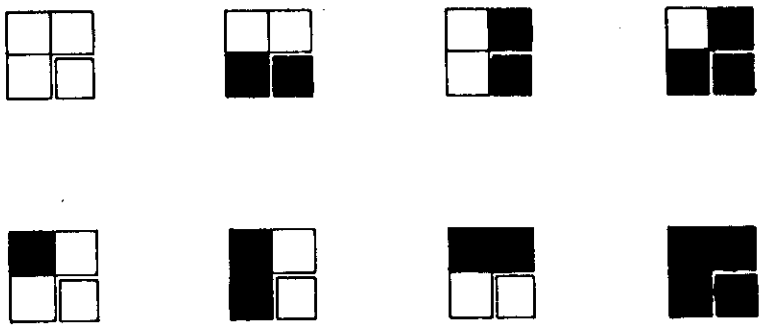
The NCI Algorithm used in the NCR prototype system at Los Gatos handles black-white images using the compression algorithm described as follows: The two step compression process proceeds as the image is first passed through a two-dimensional predictor. The output of the predictor is then passed to the run length encoder. Decompression is simply the inverse of this process.

The value of the current picture element (pel) is predicted based on the values of the immediately preceeding pel in the current scan line, and the two adjacent pels in the previous scan line. Figure 13 shows the 8 predicted values for all possible combinations of the predicting pels. The output of the predictor is compared with the value of the current pel in the original image. If they are the same, a 0 bit is passed to the run length encoder. If they are different, an error has been made, and a 1 bit is passed to the run length encoder.

The result of this process can be thought of as a new image where the 1 bits represent black and the 0 bits represent white. This is called the error image. No compression has occurred up to this point; the error image has the same number of pels as were contained in the original image. The error image is just a transformation of the original image into a form that makes the run length encoder more efficient. One result of the prediction is that the error image has less black in it than did the original.

To start the prediction process, it is assumed that there is one scan line plus one pel which are all white preceding the first pel in the original image.

The run length coder is designed to be as compatible as possible with byte oriented IBM computers and still maintain high efficiency. It uses a variable word length, but the length can take on only the values 4, 8, 12, or 16 bits. At the end of a coded page, the



Predictor Algorithm

B-F	B-F A-F	B-F A-F A-F	0-9 0-9 0-9 0-9	0-9 10-59 60-359 360-2159

Run Length Code

Fig. 13. Predictor Table and Run Length Counter Coding for NCI Algorithm