APPENDIX A

DATA PROCESSING OVERVIEW

H. S. Lahman

Introduction

The analysis of pollution incidents requires a historical knowledge of actual pollution incidents, knowledge of exposure to the opportunity for such incidents, and one or more factors correlating such incidents. Exposure data is of use in estimating the frequency of occurrence of spills; two large spills per year for a particular flag of registry may or may not be significant (or unusual) depending on the size and capacity of the fleet for that flag of registry. Exposure data may consist of a number of reasonable measures, depending on the objective of the analysis: number of port calls, days in port, fleet capacity, type of ship, location, and the like. Key factors may, of course, be anything which can be reasonably expected to correlate with spill incidents. Such key factors are usually causal in nature and, therefore, are related to volume spilled given that a spill is to occur.

The first step in the analysis was to produce a database which includes actual incidents, exposure information, and key factors. Equally important is a cross-reference system which allows correlation of exposure and key factors information to the actual spills. The major portion of this database development was the

creation of adequate cross-referencing. A secondary, but nonetheless important, consideration was to create a database which could be efficiently processed by analysis programs. Because of the disparate sources of data this presents non-trivial technical problems.

During the development of the database it was discovered that the quality of the data from certain sources had much to be desired. This required additional effort, much of it manual, to edit the data sources prior to their inclusion in the database. The errors inherent in the various sources will be discussed in detail below. Unfortunately there were no independent sources of comparison for certain key data items (notably volume spilled) so that there is no assurance that the database is completely reliable even after the editing process; it is, however, considerably better in quality than the raw data sources.

The purpose of this appendix is to describe the processing involved in establishing the database, discuss its quality, and relate the database to analysis programming referred to in other sections. All of the programs involved, with the exception of the standard IBM sort routine IGHRCOØØ were programmed in PL/I. Development was done on an IBM 370/158 system which included the WYLBUR timesharing editor and preprocessor. In the following sections the reader will note an unusual number of programs designed to list intermediate files. This was not simply a debugging mechanism; though not shown in the flow charts there was

typically a manual editing of the produced files, via WYLBUR, based on such listing. These editing steps eliminated duplicate records, obvious coding errors, records with internal inconsistencies, and the like.

Primary data sources

From the standpoint of computer manipulation, there were three primary data sources: the U.S. Coast Guard Pollution Incident Reporting System (PIRS), the U.S. Geological Survey LPR1Ø lease block information system, and the Martingale Master Vessel File (MVF). The PIRS file dealt with actual pollution incidents. The LPR1Ø data provided platform production data for Gulf Coast OCS leases. The MVF provided data on the makeup of the world tanker fleet. All three of these files were in computer-compatible format. In the actual analysis, we also used the USGS accident, structures, and pipeline files. However, these files were not obtained in computer-compatible form due to file management difficulties. In addition, there were a number of other published data sources which provided useful information. Most notable among these other sources was the U.S. Army Corps of Engineers (USACE) port call data.

A number of other data sources were considered (e.g., the Louisiana state spill data) but, generally speaking, these sources were unsatisfactory due to lack of coverage (Louisiana destroys its data after six months), inherent low quality, or incompatibility with other data sources.

PIRS data

The USCG PIRS file provides comprehensive information on all spills occurring in U.S. waters under USCG jurisdiction. In addition to the computer-compatible PIRS file, there are, in theory, written reports available. However, in checking with the Eighth Coast Guard District headquarters, we found the written records had been destroyed for the period 1973-1975. Thus, there is no back-up for the PIRS OCS spillage data.

The PIRS file contains, in theory, abundant information concerning each spill reported. Aside from location, date, and volume spilled, the file contains a variety of data related to potential causal factors (e.g., ship size and type, type of operation during the incident, primary cause, secondary cause, and the like) and environment (e.g., weather conditions, water body type, etc.). Generally this information is in the form of codes representing specific classifications. Such information is potentially extremely useful for correlating volume spilled with key factors.

Sadly, this source is plagued by an apparent lack of quality control. Martingale has manually checked a number of specific data fields where there were independent sources for the data. The conclusion is that for at least some of the coded fields there is about a 5%-20% chance that there will be a coding error (e.g., vessel size, vessel type). Whether these are recording or keypunch errors is

problematical; the fact is that quality control on this file is far below commercial standards. Assuming the errors are due to translations from literal description to code, it might be expected that the frequency of errors might be diminished for numeric data.

Coding errors present potentially significant analysis errors when dealing with pollution incidents due to the fact that catastrophic events such as collisions are relatively rare but highly significant. Thus, a single coding error on the cause of an incident (say, personnel error rather than equipment failure) involving a very large spill could have a large effect on the volume distribution for the relevant causes.

In addition, there are problems with the coding/ recording system itself. Some data fields, notably wind and current direction, are rarely assigned values, thus raising the possibility that any derived statistics might be highly biased. Some codes, such as Water Body Type, proved virtually useless for analytic purposes, due mainly to the ambiguity inherent in the wording used to define water bodies (i.e. "bay" could mean an enormous water body, as with the Bay of Biscay or Hudson's Bay, or an inlet of small dimension). In other codes there appears to be significant overlap in definitions. For example, the operation type field contained a number of activities that could be carried out simultaneously, e.g. underway and pumping bilges. Again, this sloppiness causes this portion of the data to be virtually useless.

There are also problems with aggregations: there are peaks in the spill volume histograms at 10, 100, 1000 etc. gallons, reflecting interpretive roundoff in the estimates ("that's about two barrels, which is about 100 gallons . . ."). There are also very sharp peaks in the current and wind speed histograms at 5, 10, 15, and 20 knots, again reflecting interpretive roundoff.

During the editing process Martingale corrected those data fields which were in error relative to independent sources. In a number of other cases records were eliminated due to internal inconsistency or obvious errors. Since this elimination process potentially introduces bias to the analysis, all such eliminations were carefully examined. Typically, if the spill was of any significant volume it was found that the record was for an anticipated spill and not an actual occurrence. We feel that no harm was done to the analysis by this culling process.

LPR1Ø data

The U.S. Geological Survey LPR1Ø system contains comprehensive data on OCS (federal) leasing and production. The LPR1Ø file contains data on location, production, and leasing over the life of all OCS leases. For this report the leasing data was ignored. Production data is broken down by year for crude, condensate, gas, and miscellaneous. Location information is in the form of area, block, and lease numbers.

Though it was of considerably better quality than the PIRS data, there were still some quality-control problems with this data. In routine checking of platform location for latitude and longitude we discovered that producing platforms and leases appeared on published maps but not in LPR1Ø. There were also some technical problems related to out-of-date file format documentation and unwieldy file structure. Such technical problems were overcome with additional processing and had no effect on the analysis.

Martingale had no means of checking accuracy of production data against independent data sources, so these values were taken on faith. It should be noted, however, that there were inconsistencies between PIRS and LPR1Ø in that PIRS records production-related spills, in some instances, when LPR1Ø indicates no production in that year.

MVF data

The Martingale Master Vessel File contains data on the world oceangoing fleet of tankers, bulk carriers, and combined carriers. This data was compiled from a variety of sources, most notably the American Bureau of Shipping, Clarkson's Register, the U.S. Maritime Administration, and Fairplay. Considerable effort was put into developing this file and reconciling inconsistencies among the various sources. For the data fields essential to this report (i.e., call sign, flag of registry, ship type, and capacity) the file can be regarded as authoritative for those ships it contains (at least 85% of the entire world fleet).

Other data sources

Some exposure data was based on the USACE published data. This data was primarily used to extract traffic data through port calls. No judgment was made on the quality of this data, although other researchers have found it to be partially incomplete (Fricke, 1977). The USACE maintains a computer-compatible data base which presumably contains much more information that that published. If this data were merged with the Census Bureau AE350 and AE750 data, fairly detailed traffic estimation could be made. In particular, to establish traffic over specific routes (e.g., coastal waterways) it would be possible to systematically interpret embarkation/destination data for each port call. Such analysis was considered, but time and dollar constraints put it far beyond the scope of this study.

In order to cross-reference PIRS and LPR1Ø data it was necessary to translate latitude/longitude to area/block, and vice versa. This was done manually from lease maps published by USGS. These maps presumably meet conventional mapping standards and, therefore, are accurate to within the one minute necessary for cross-referencing. There was some difficulty because of the rather bizarre block number schemes employed and the ambiguity of area boundaries. Also, some block numbers were missing from the maps. To attempt to reconcile these difficulties, maps prepared by the Offshore Publishing Company for the Oil and Gas Journal were employed. These proved to be more up-to-date than the USGS maps, but the scale was too large to be useful for pinpointing latitude/longitude, so they were used primarily to resolve ambiguities in the USGS charts.

Database development overview

The purpose of developing the database was to generate a group of files which could be conveniently utilized on a repeated basis by the various analysis programs. The primary data sources were in a variety of formats, some not even computer-compatible. Also, the primary sources generally contained much more information than was necessary for the study, which meant added processing overhead (e.g., on-line storage costs). Thus, before any analysis could be performed, the primary data sources had to be reduced to a usable and efficient form.

The first major effort was to create extracts of each of the primary data sources which contained only essential information in a format readily processed by the analysis programs. Such extracts reduced storage and processing costs as well as reconciling format incompatibilities. A byproduct of this task was the production of master file listings. These listings allowed evaluation of data quality and detailed planning for further processing and analysis.

The second major effort was cross-referencing among the various primary data sources. Between PIRS and LPR1Ø this required a cross-reference between area/block and latitude/longitude. Between PIRS and the MVF the cross-reference involved call sign or official number.*

*Official number is not contained in the MVF but is readily correlated manually from indirect sources such as the American Bureau of Shipping Register.

Both these cross-references were developed manually for the most part. Once the cross-references were established, data fields were added to the primary source extracts which provided a means of mutually keying these files.

The third effort was to generate from the extracts of the primary sources a set of files for use with specific analysis programs. Since the analysis programs were executed many times, it was desirable to format these input files in an efficient manner. Only data necessary to the specific program was used and typically the files were sorted in a manner which yields best efficiency. The overview schematic of these efforts is shown in Figure A.1.

Another effort, which overlapped the previous three, was to edit the data. At each phase of data reduction and during analysis the results were examined for internal data inconsistencies and errors. In addition, comparison was made with independent data sources to detect inconsistencies. Where ambiguities were found, they were either corrected or the records deleted from the database.

The relationship of the database to the various study tasks is shown in Table A.1. PIRS data was employed for the tanker analysis to generate distributions and histograms on volume spilled, usually through the use of DISTRIB, POSTLIK, and other programs. Extensive regression analysis was performed on the PIRS spill data and the LPR1Ø lease production data in an attempt to establish correlation. USACE data and a summary of the fleet (MVF) listing provided information for

FIGURE A.1
SPILL DATABASE OVERVIEW

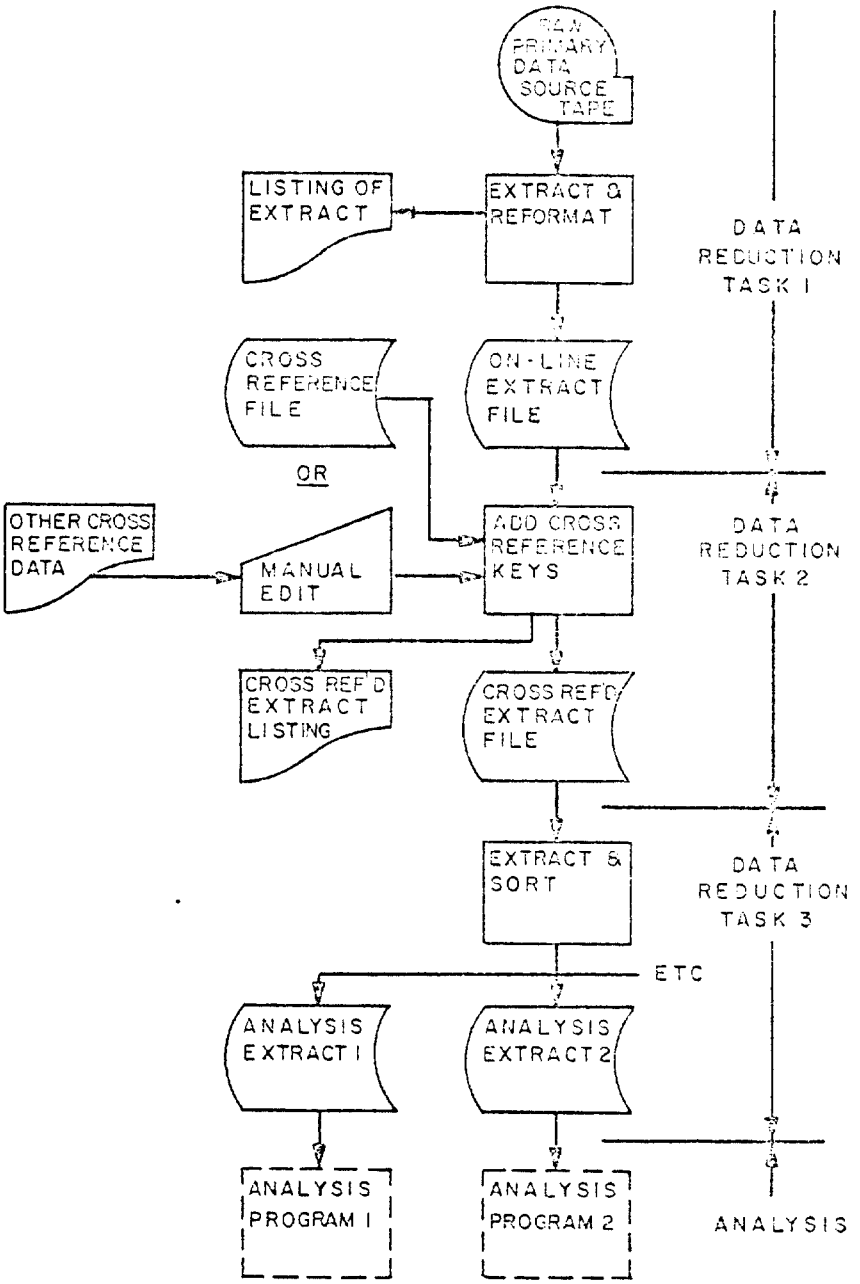


TABLE A.1
RELATIONSHIP OF DATABASE TO STUDY TASKS

Task	File/Listing	Program/Operation
1	PIRS production spills LPR1Ø production extract Independent data sources	DISTRIB, etc. Volume distributions
2 (tankers)	PIRS tanker spill file Summary of fleet by flag USACE data	DISTRIB, etc. Volume distributions Ad hoc frequency of occurrence
2 (platforms)	PIRS production spills LPR1Ø production extract USGS accident, structure and pipeline files	REGLPR--regression analysis DISTRIB, etc. volume distributions Ad hoc frequency of occurrence
3	Various file listings Independent data sources	Quality analysis

frequency-of-occurrence analysis of tanker spills. The spill volume analysis for offshore platforms and pipelines was based primarily on the USGS data. The various listings produced during the database development provided the basis, together with independent published sources, for analysis of overall quality and suggestions for future revisions in reporting/recording practices.

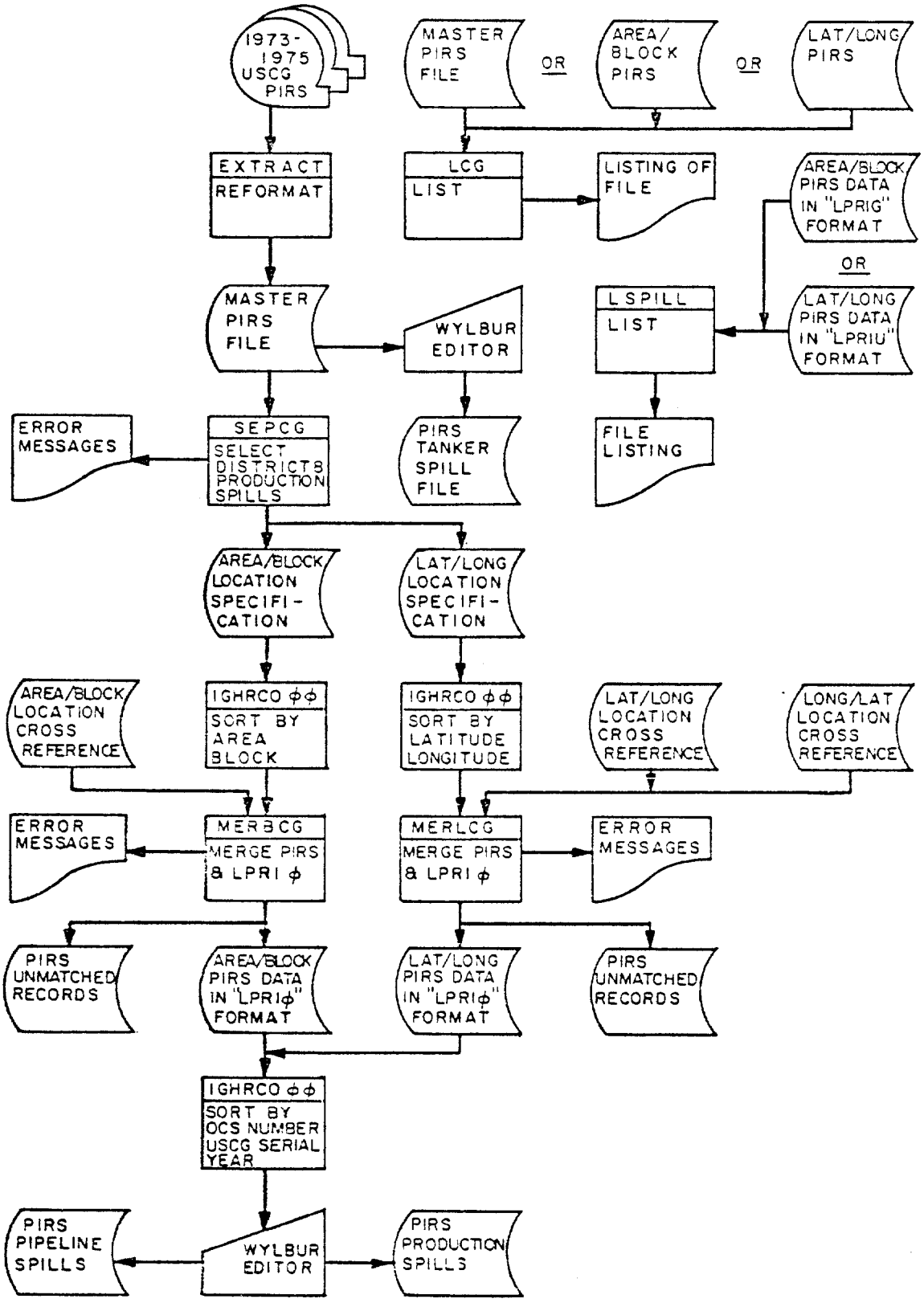
Detailed database development

PIRS processing

One computer tape containing files for each of the years from 1973 through 1975 was obtained from the U.S. Coast Guard. Within the three files, each logical record represents data for a single reported spill incident. This tape contains all reported spills, in theory, so the first task was to extract those spills of interest to this study (e.g. tanker spills and Gulf Coast platform spills). Because the PIRS records contained a number of data items irrelevant to the analysis (e.g. information on fines) it was also desirable to reformat the records extracted so that they contained only data items relevant to the study--or for statistical pollution analysis in general.

The program EXTRACT (Figure A.2) was employed for these purposes. EXTRACT identified the ten categories of data of interest. One of these pertained to the Gulf Coast OCS spillage. The other nine related to tanker or tank-barge spillage in the rest of the U.S. Each category was signified by a character in Column 80 of the reformatted record. This

FIGURE A.2
PIRS PROCESSING



allowed ready editing and manipulation of specific data groups via WYLBUR.

Production spill extraction was somewhat more complicated due to the number of possible code values related to production spills and the alternative modes of location specifications employed by PIRS (i.e., either area/block or latitude/longitude. This feature of the data may have been corrected with the most recent PIRS coding instruction, which calls for latitude and longitude for OCS production platforms.) The program SEPCG was employed to extract all USCG District Eight (Gulf Coast) production spills and place them in two files, depending on the location specification. These two files, in turn, were sorted by location in anticipation of merging with a location cross-referenced file generated from LPR1Ø production data. The program LCG was a utility program which listed the various extract files in a convenient format for manual editing.

At this point in the PIRS processing the various file listings were examined for errors. (EXTRACT and SEPCG also flagged certain records for obvious errors such as alphabetic characters in pure numeric fields and lack of volume or location data.) Where possible the data was checked against independent sources of information. Errors were corrected in individual records through the use of the WYLBUR editor. When errors were found which could not be corrected the entire spill record was deleted from the file.

The next task was to match location specification information for production spills so that spill records were defined by USGS OCS lease number for subsequent matching with production data. At the same time the spill records were reformatted to a quasi-LPR1Ø format in case it might be desirable to incorporate these records with the LPR1Ø database in the future. This was accomplished by the programs MERBCG and MERLCG. MERBCG dealt with spill records where location was defined by area/block codes, while MERLCG dealt with latitude/longitude specifications. A location cross-reference file containing area/block, equivalent latitude/longitude, and OCS lease number was used for this purpose.

For MERLCG two identical cross-reference files were used, one sorted by latitude then longitude, while the other was sorted by longitude then latitude. This was necessary because nearly all OCS leases cover more than a minute of latitude or longitude so that there could be some ambiguity in such specification. The MERLCG logic examined latitude for an exact match and the longitude to a match within one minute. If no match was found, the logic sought an exact match in longitude and a latitude match within one minute. The case where both latitude and longitude were off by one minute from that designated in the cross-reference file was not examined.

MERBCG and MERLCG produced reformatted spill files keyed by OCS number. The programs also flagged unmatched

records (i.e., those USCG designated production spills where there was no active USGS lease). Roughly 20% of the PIRS production spill records could not be matched to OCS lease locations. Manual checking revealed that some (perhaps 5% or so) were due to missing data in the LPR1Ø database (i.e., commercial sources show active platforms where LPR1Ø has no active lease). Some of the mismatches of latitude/longitude specifications were due to the failure of MERLCG to check both latitude and longitude discrepancies of one minute.

The next step was to merge the production extract files and sort by OCS lease number, USCG serial number, and year. These files could be listed by the utility program LSPILL in a format convenient for edit verifications. For use in regression analysis it was desirable to remove these spills that had no corresponding entry in the LPR1Ø file and condense the records into a single record format (the extract files contained two 80-character records per spill). This was done via the WYLBUR editor.

LPR1Ø processing

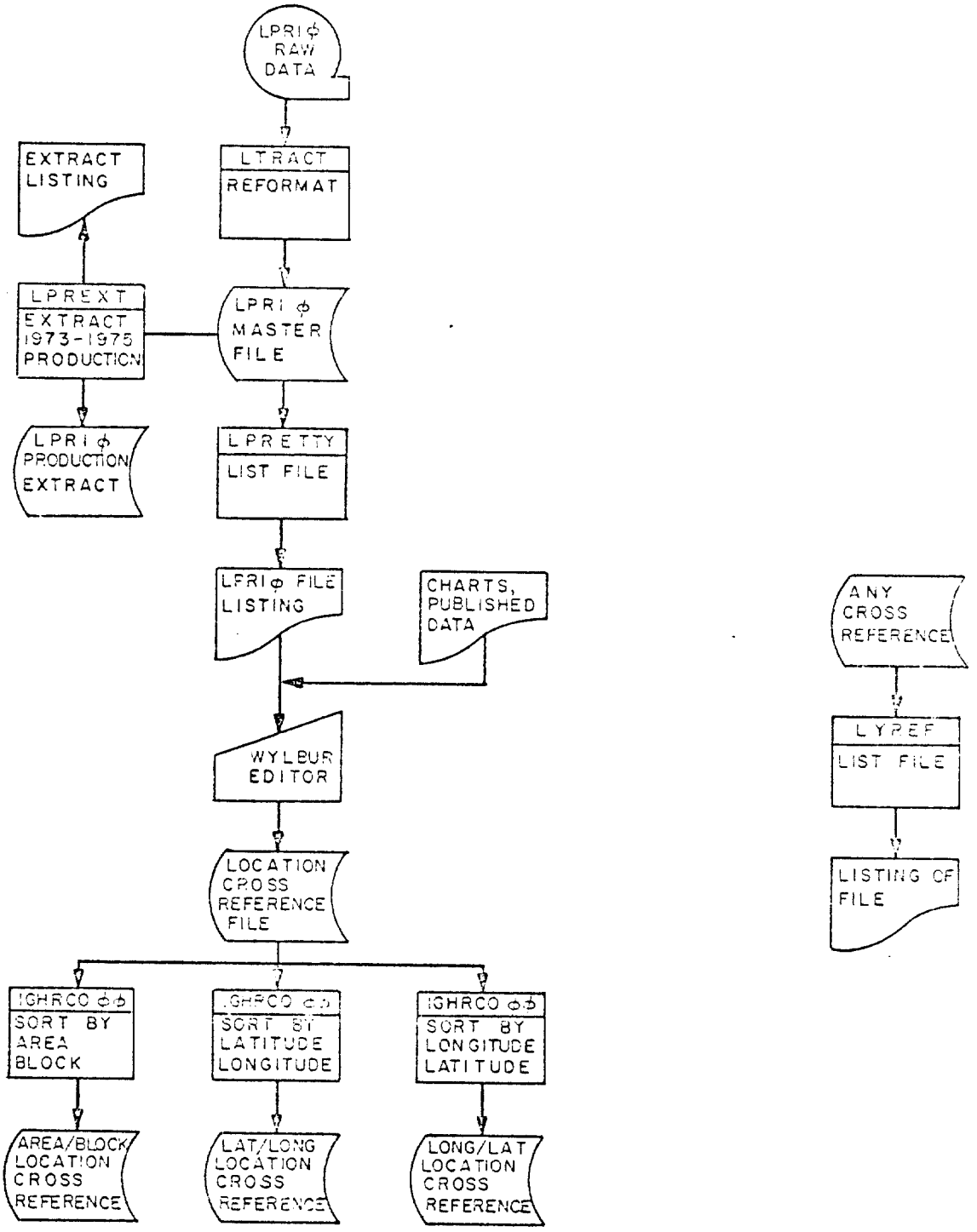
The LPR1Ø database provided from USGS was a computer tape. There was no documentation concerning the tape contents and the first task was to find the database among the extraneous (and unrequested) files contained on the tape. This necessitated a series of tape dumps in hexadecimal format. The LPR1Ø database contains multiple

records for each OCS lease. Much of the information concerns financial parameters (e.g., lease payments, royalties, etc.) which were of no interest to this study.

The first task was to extract and reformat the Gulf Coast information into a useable file. This was accomplished with the program LTRACT (Figure A.3) which created a master LPR1Ø disk file. The program LPREXT was then employed to extract and condense 1973-1975 production data for use with the regression routines. LPREXT also listed the production extract.

The program LPRETTY was employed to list the master LPR1Ø file for editing and reference purposes. This listing, together with charts and other published data, was employed to create a location cross-reference file containing OCS number, area/block (with state due to duplications), and latitude/longitude. The location cross-reference file was created by direct data entry through the WYLBUR editor. During this editing process it was found that a number of active platforms, perhaps 5%-10% of the total, were not found in the LPR1Ø database. Where possible these were added to the location cross-reference file. The location cross-reference file was then sorted in three different ways for subsequent merging with PIRS data: by area then block; by latitude then longitude; and by longitude then latitude. The program LXREF was employed to list the file in useful format for editing.

FIGURE A.3
LPR 10 PROCESSING



MVF processing

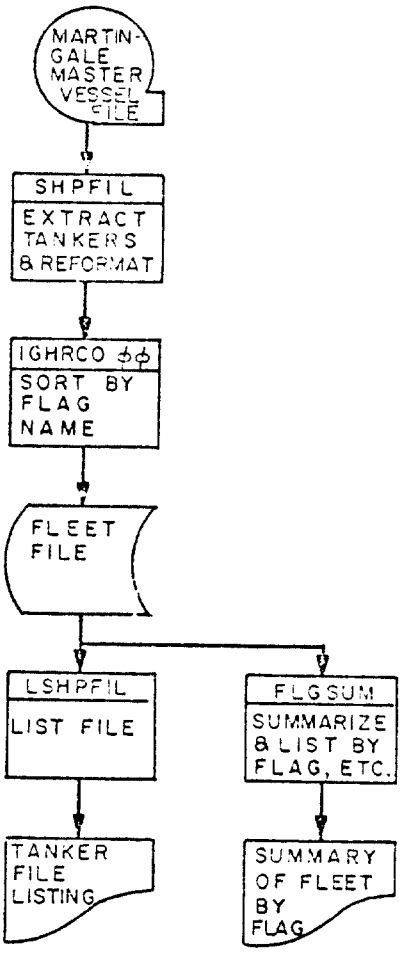
The Martingale Master Vessel File contains a large amount of data for each ship in the world tanker, dry bulk, and combined carrier fleet. This data was synthesized from several commercial sources and is quite reliable for purposes of this study. The file does not contain all ships in the fleet; perhaps 10%-15% of the fleet, in number of ships, is not in the file. Those missing are generally small coasters and the like; total fleet capacity of the file agrees quite well with independent estimates.

Processing of the MVF involved extraction of tankers and culling of information (e.g., speed) which was not necessary to the study. This was done through the program SHPFIL (Figure A.4). The extract file was then sorted by flag of registry and name. The program LSHPFIL listed the extract in useful format for referencing against other sources. The program FLGSUM was employed to summarize the fleet by flag category, capacity, and the like. The listing produced by FLGSUM was the basis for exposure data in tanker frequency-of-occurrence analysis.

Relationship to programming and analysis

In various forms the database provided a basis for two functions: analysis of frequencyof occurrence of spills and analysis of volume spilled given the occurrence of a spill. Generally speaking analysis of volume spilled was done analytically on the computer while analysis of frequency of occurrence was done manually in an ad hoc manner. Analysis

FIGURE A.4
MVF PROCESSING



of volume spilled was done by generating histograms and sufficient statistics with the program DISTRIB (Figure A.5). Essentially DISTRIB provided selection criteria (e.g., selecting all records with spills related to tankers) and summations on number of spills for various volume categories and other parameters.

The sufficient statistics provided by DISTRIB were then used as input to BAYLOG, POSTLIK, and POSTVOL. BAYLOG generates the posterior probability on the next spill for a lognormal distribution; POSTLIK, the posterior probability for hypothesis testing; and POSTVOL, the posterior distribution on volume spilled for gamma and inverse gamma distribution.

Though frequency-of-occurrence estimates were made manually, there were two important exceptions. For the failure/spill model a Monte Carlo process was employed to simulate the events producing spills. This was done with the programs WHACKO and CAMILLE and utilized independent published data for the estimation of key parameters. The second exception came with regard to analysis of frequency of occurrence of platform spills. In this case regression analysis through the program REGLPR and its subroutine STAT1 was performed in an attempt to explain frequency of occurrence through production rates and platform operator.

FIGURE A.5
DATABASE / PROGRAMMING