CSF

DATA ERROR ANALYSIS--

LAND USE STATISTICS

362 - 67

AUTHOR: K. CRANE

COMMUNITY SYSTEMS FOUNDATION

TABLE OF CONTENTS

INTRODUCTION

The primary objective of this project was to process specific information for the Pilot Study Area for use by the City both in the analysis of that area and as a means of evaluating information outputs to be used in the analysis of the entire city.

It was recognized that because of both the large volume of data that will ultimately be collected by the City (400,000 to 500,000 records) and the complexity of this data (53 fields per record), it would be impractical to process every possible combination of city information even if it could be done in one pass. It would be equally impractical to attempt to analyze all possible combinations of this information.

Furthermore, because there are numerous ways to display a given piece of information, it is desirable to know which kinds of displays will be most useful in analyzing city information.

Therefore, as a means for the City Planning Department to evaluate which kinds of information would be most useful in the analysis of the entire city as well as the most effective manner in which to display this information, a Pilot Study Area was defined within the city and selected information for this area was retrieved and displayed in the form of tabulations, bar charts (arrays), bar graphs (distributions), and maps.

A secondary project objective was to document the problems encountered in processing the Pilot Study Area information so that they may be corrected or circumvented when processing data for the entire city.

Most of these problems were noted in the initial Community Systems Foundation report entitled "Logic for Producing Land Use Statistical Reports." The remainder of this report will be concerned with further discussions of these problems.

COMMUNICATION PROBLEMS

Most of the difficulties in using information storage and retrieval systems result from communication break-downs between the people (1) who specified the data contained within the system (2) who collected and stored the data (3) who will specify the information to be retrieved from the data and (4) who will retrieve the specified information.

The people who specify data are concerned with general information needs, data availability and data cost. People who collect and store data are concerned with the mechanics and practicalities of data collection methods, formats, organization and indexing. People who specify information have a host of questions - and hopefully some of them have answers buried within the data. People who retrieve information are concerned with performing arithmetic and logical operations on specific data fields within selected records - with a minimum of computer time. This diversity of principal concern is then complicated by diversity of technical backgrounds, personnel turnover and inter-departmental communication channels.

The objective of the report "Logic for Producing Land Use Statistical Reports" was to eliminate communication problems between the planners who specify information and the computer technicians who retrieve this information from the data records. However, despite this effort,two problems still arose. They were as follows:

1. Undeveloped land was imprecisely specified as all parcels with a first-digit land use code of "9". It was later learned that these parcels could include some developed, but vacant, space uses (2-digit code "94") as well as some land under development (2-digit code "95"). Much of the retrieval logic and some retrieval programs had to be modified to accomodate the 2-digit, "94", codes. The "95" codes were left as "undeveloped" inasmuch as there was no more appropriate category for that land use.

2. "Artificial" census block numbers had not been explained. Thus, the complete lack of data for certain blocks (tract 16, blocks 35, 40 & 41; tract 17, blocks 41, 42, 43, & 44; tract 18, block 57), and the unexpected appearance of other blocks (tract 16, block 38; tract 17, blocks 8 & 9) came as a surprise. Fortunately, this communication break-down only affected the preparation of mapping programs.

The majority of the problems encountered on this project were discovered while developing the retrieval program logic (i.e. - prior to and during, the preparation of the report - "Logic for Producing Land Use Statistical Reports"). It turned out that many initially desired outputs could not be produced due to limitations in the data record or file structure. These problems are as follows:

1. "Tax exempt land acreage by land use code" could not be produced because parcel tax status is not in the data record. Parcel tax status was on the original data item list but was dropped somewhere along the line.
2. "Total first floor space" in any multiple story building cannot be accurately determined whenever the establishment (s) within that building occupy both the first floor and additional floors. The "floor space" field contains the total floor space for each establishment. When this floor space is on more than one floor, the "floor level" field indicates the predominent floor.
3. The hierarchy of parcel level, space use level, and establishment level records in the data file created the bulk of the retrieval problems.

   A. Although each record is an "establishment" record and all have the same format, only the predominent establishment within a space use has data in certain space use level fields (e.g., predominent land use code-

space use, space use zoning, number of establishments, building address information, and building characteristics information). Likewise, only the predominent space use level establishment within a parcel has data in certain parcel level fields (e.g., parcel area, average parcel width, average parcel depth, parcel frontage, predominent land use code-parcel, niusances, number of space uses). Because of this lack of space use and parcel data in every establishment record, any sorting or processing sequence that separates corresponding parcel, space use and establishment level records also hamstrings the retrieval of information correlating the three levels. Such was the case in the outputs described in contract appendix A, table group D-1 and D-2. The sort sequence for these tables had to be building condition within standard land use code. Hence, "building condition", which was space use level data, had to be transferred to corresponding establishment level records prior to processing. Also, "number of floors" had to be transferred in order to estimate "total first floor space".

B. Parcel data, in particular parcel area, parcel width, and parcel valuations (when they become available) are not apportioned to the space uses or establishments within the parcel. Hence, it is impossible to accurately reference this parcel data in terms of space use level or establishment level data. For example, parcel area cannot be accurately referenced by 4-digit standard land use code as this code is establishment level data. Neither can parcel area be accurately referenced by zoning category as zoning is space use level data. Similarly, space use data, in particular number of dwelling units and number of establishments, are not apportioned to the establishments within the space use. Therefore, it is impossible to accurately reference this space use data in terms of establishment level data. For example, number of dwelling units

cannot be accurately referenced by 4-digit land use code. This constraint had to be faced in most of the Pilot Area outputs and are noted as they occurred in the "Logic" report. For the most part, approximate referencing between record levels was accomplished by using the data from the first (predominent) establishment encountered within a parcel or space use. Some information was not produced at all due to this constraint - namely:

1. Total developed acreage, excluding easements, by census blocks.
2. Average parcel area by building condition and residential land use code.
3. Average parcel area per dwelling unit by building condition and residential land use code.
4. Floor area divided by parcel area by building condition and land use code.
5. First floor area divided by parcel area by building condition and land use code.
6. Floor area divided by parcel area by building condition and land use code.

It is likely that these problems stemmed from a communications breakdown between the people who specified, collected and stored the data and the Planners who specified the information to be retrieved from the data.

If communications between these groups had been clear and complete, either the data file would have been constructed to facilitate (and enable) retrieval of all requested information, or, the planners would have had a better understanding of the files limitations and would not have requested information that could not be retrieved.

DATA PROBLEMS

Except for errors in tape file characteristics (record formats, sort sequence, blocking, recording density, tape lables, etc.), the majority of data errors will never be discovered, once they are on tape, without the use of one or more editing programs.

There are two basic types of record editing programs - one type checks data field specifications (e.g., checking for alpha characters in an all-numeric field, or checking the range of values in a limited numeric field), the other type makes logical checks within each record or between records (e.g., summing fields A, B and C and comparing for an equal with field D).

The field specification edit is useful for discovering gross errors quickly. If this type of edit is not performed, most of these errors will probably be discovered (painfully) during the information retrieval process. The Pilot Study Area data had not been edited for field specifications and the following errors were found:

1. The tract number on 175 records was punched "2J" instead of "24". This error was discovered after the output described in appendix A, table group A, had been produced. All other output represents corrected data.
2. The block number for artificial block 57 in tract 18 was not punched for all records in this block (7 records). This error was handled as described above.
3. A few scattered parcel level records had no parcel area data. No correction attempt was made.

The usefulness of the record logic edit depends upon whether or not there are identifiable relationships between data fields within a record or between records. Naturally, if all data fields were independent, this edit would be

worthless. However, this is not the case with the Kansas City file since there are definite relationships between parcel, space use and establishment level records as well as within each parcel level record.

One error that could have been found with a logical edit was discovered while calculating parcel area per dwelling unit for parcels with certain zoning and land use codes (Contract appendix C, data item 10). A few scattered parcel level records had no "number of dwelling units" data although they did have residential land use codes (1111 and 1131). This would mean either the number of dwelling units field or the land use code field were in error for those records.

This kind of error would not have been caught in a field specification edit since, considered independently, "blank" is a legitimate code for number of dwelling units and both 1111and 1131 are legitimate land use codes.

Other errors or ommissions in the Pilot Area data were discovered or acknowledged by the Kansas City Planning Department and related to Community Systems Foundation. They are as follows:

1. The City plan block number for 18 parcels within tract 16, block 18, was incorrect. This error did not affect the desired information output and was not corrected.
2. Three records were missing from the pilot area data. They were not supplied and are not reflected in the Pilot Area information outputs.
3. Assessed valuation data had not been included in the data shipped to Community Systems Foundation for processing. Because this data was not later supplied, information relative to this data could not be produced.

FILE MAINTANENCE

The errors within the Pilot Area data involved both individual records (e.g., the missing parcel area data in a few records) and groups of records (e.g., the records with the mispunched tract number). It is likely that the entire city file also contains similar random and consistent errors.

To insure that these errors can be corrected easily and also to insure that the data can be kept current, file maintanence programs are required that contain flexible record selection and field replacement features. Typical file maintanence programs select records based upon fixed identification fields and then replace the contents of the entire record with corrected or updated data even though most of the record does not change.

Correction of consistent errors can often be simplified if record selection fields are variable depending upon the problem at hand. Also, replacing selected field data rather than all data, minimizes the possibility of creating additional errors in the maintanence process.

Updating of data has been simplified by the incorporation of many identification and location fields into the record format (22 of the 53 fields identify and/or locate each record) and by using small units of data (establishments). Thus, it is likely that the data in the cities file will be compatible with that in the files of other governmental units or agencies. However, it is unlikely that the record formats within these files are identical to the cities' format. Hence, updating will have to be on a field by field replacement basis.

RETRIEVING CITY-WIDE INFORMATION

Much of the information requested from the Pilot Study Area file required complex processing due to the overlapping conditions for record selection (e.g., conditions on space use number, establishment number, building condition, and land use code had to be met prior to selection). In addition, record processing sequence was crucial due to the file structure (i.e., the different record levels).

If the kinds of information retrieved from the Pilot Area file are representative of the information desired from the city file, then it is highly doubtful that any one utility retrieval program could produce all of the results desired.

However, rather than creating a battery of utility programs to produce generalized sub-sets of information, it would probably be more practical to have one utility program backed up by specialized programs which produce specific, more complex, information. This course of action assumes that information needs have been carefully thought out and specified in light of the utility program's limitations.