HANDLING SMALL AREA DATA WITH COMPUTERS

Richard S. Hanel
Vice President and Manager
Urban Statistical Div.
R. L. Polk & Co.
551 5th Ave.
New York, N.Y. 10009

"It's no trick at all for the computer to generate a million or so statistics for one block in a medium-sized city. The big job is to figure out what to do with a million numbers — how to work with them and what to conclude from them."

This article deals with today's capabilities for handling small area data with computers.

We will be dealing with three main points. First, a summary of eight important computer capabilities which are ready to be put to work right now. Second, some examples of practical, results-producing applications of those capabilities. And third, a few words of caution.

Our emphasis is on what can be done right now — *today* — in terms of what we consider to be an exciting new dimension in statistics — the dimension of space — which includes physical location and geographic relationship.

Credentials

First, though, just a few words to position the Polk Company and give our credentials for discussing computers and small area data. Our company is the "official scorekeeper" for the automotive industry. Each year, we process nearly 100 million car and truck registrations, in order to give the auto companies the detailed small-area statistics they need to keep track of all the cars and trucks which are sold and are on the road. Our experience in using computers to code and summarize vehicle registrations by small-area — such as census tract and dealer trading zones — goes back to the early 1960's.

Another important part of our business is the publication of City Directories. Each year, Polk interviewers go door-to-door in some 7,000 communities across the United States to gather the information which is printed in our City Directories. All told, we knock on the doors of some 24 million households and 3½ million businesses each and every year, as we take our annual City Directory Census of well over half the urban population of the U. S.

In the very early stages of programming our computers to sort and print the Directory information, we found that it would also be possible to turn the interviews into statistics.

Based on a talk at the Federal Statistics Users Conference, New York, N.Y., September 12, 1968.

So it is that for the last 5 years or so, we have been deep in the business of preparing address coding guides, coding and summarizing data for areas as small as a block, and designing the kinds of computer output with which the numbers could best be put to work.

20,000 Miles of Computer Tape

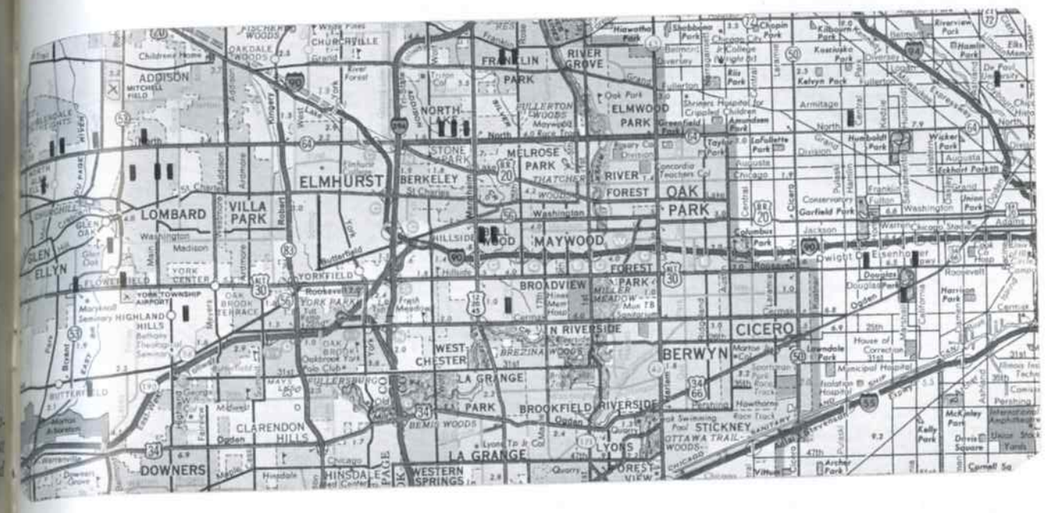
Company-wide, we're using close to $10 million worth of third-generation computers — and some 20,000 miles of computer tape — to store and process just the current information in our files. And wherever you look in the Company, an important part of our computer activity is devoted to sorting and summarizing data by some kind of small area — be it street, block, tract, planning area or zip zone.

Now then, what have we learned in the last 5 years that might be of interest and use? In general, three things:

— First, there exists right today, plenty of computer capability for dealing very effectively and at very low cost with small-area data. We have all the equipment we need right now to plow the small-area fields thoroughly and efficiently.

— Our second main observation is that the people who are starting to put small-area capability to work, in even the simplest and most straightforward ways, are discovering that the computer is adding a whole new dimension to the meaning and use of statistics — the tremendously important dimension which includes physical location, density and geographic relationship. We now have the ability to track statistical change through space just as effectively as we've learned to track change through time.

— Third, we find that the computer is turning out to be a stern and demanding taskmaster which is setting a fast and tough pace for those who are pioneering in the use of small-area data. For example, it's no trick at all for the computer to generate a million or so block statistics for a medium-sized city. The big job is to figure out what to do with a million numbers, how to work with them and what to conclude from them, who's going to use them and how.



These are our three general conclusions. Now let's get specific — first, in terms of a look at some of the computer capabilities that exist right today for dealing with small-area data.

Computer Capabilities

Our starting point is the job of translating tens of thousands — or even millions — of street addresses into their equivalent small-area identifiers, such as block, tract, planning area or Zip Zone number. This job of *geo-coding* can be done by computer at the rate of 5,000 to 6,000 addresses a minute — roughly 100 per second — using a sorting table known as an address coding guide. The U.S. Census is preparing coding guides for many cities for use with the 1970 data. The main point here is that for many purposes there's no need to wait for the work that's being done for the '70 census — it's simple and inexpensive to set up your own customized, computerized coding guide which can go to work on your own data right away.

Once your records are geo-coded, the ability of the computer to store, select and summarize huge masses of data at blinding speed takes over. This capability for fast, low cost *mass-processing* is the foundation on which the whole small-area data business is starting to be developed. When you stop to think of it, it's only in the last few years, with computers, that we've learned to deal effectively with the reams of block data that have been available ever since the 1960 Census.

Flexibility

A third capability — and an outstanding characteristic of the computer information systems which are being developed and used these days — is the almost incredible *flexibility* with which computers can select and retrieve and manipulate the data in the files. A number of readily-available programs are making the design and production of statistical tables, summarizing data by small-area, a relatively uncomplicated and routine matter.

Another important capability of computers in dealing with small-area data is their tremendous power for developing valuable *by-products* as they perform their basic sorting and counting operations. For example, as we summarize our automobile registration counts by small-area, we are simultaneously calculating and summarizing the probable number of miles those cars will travel in the next year, and how much gas they'll burn up in the process. In order to do this, the computer bounces each registration against a probability table which takes into account the make and age of the car, the number of cylinders, whether it's registered in the name of a male or female driver, and how many cars in total are registered at that particular household.

Another capability that we're learning to work with is the way the computer can use address or small-area designations — such as block or tract number — as a common denominator for *merging information* from any number of diverse sources, both internal and external.

An automobile insurance company, for example, can take an internal count of its policyholders by ZIP area and compare those figures with an external count of the number of car-owning households by ZIP. By calculating the percentage of policies to households for each of the 35,000 ZIP areas in the country, you have the beginnings of a very effective, small-area sales management and control tool. Incidentally, one of our computers is capable of performing the 35,000 percentage calculations in a little under 5 seconds. Printing out the 35,000 answers, including ZIP number and a six or eight digit percentage for each, takes a little longer — about 5 minutes.

With 35,000 numbers to deal with, we come now to another — and one of the most intriguing — new computer capabilities. At one stage of the game, we thought that well-designed statistical tables, itemizing small-area counts and subtotals, and complete with a wide variety of summaries, comparisons, percentages and ratios, would be just what the data users and decision-makers were waiting for.

As you might guess, it didn't turn out that way. For one thing, in most cases, there were — by definition — just too many numbers. 35,000 percentages, for example, are a bit much for easy review and decision-making. And more importantly, it's difficult for a statistical table to do any kind of justice to the discovery and display of geographic relationships in the data.

Graphics

We're beginning to find some answers to this kind of problem in the technique of *computer graphics*, which represents one of the most important new computer capabilities in dealing with large volumes of small-area data.

Quite possibly you have seen examples of maps which have been printed on computers. At one end of the line are the relatively simple profile maps in which specific areas — such as blocks or tracts — are clearly outlined and shaded in with a regular computer printer to indicate various levels of value in the data. At the far end of the range are the very precise and elaborate line tracings which are done on complex and expensive special equipment involving rotating drums, photoelectric cells, highly sensitive film and all manner of weird and wonderful devices.

One of the most interesting and advanced techniques is one which summarizes geographic data by calculating and printing the equivalent of contour or isotherm lines, and has the further ability to rotate the contours and look at them in three dimensions from the top or the side, or any angle in between.

It used to be that maps found their greatest use as the end product or display piece with which decisions already arrived at were justified or explained. After all, it used to take hours or even days to prepare outlines, locate and plot data, stick in colored pins and fill in the shadings for a map dealing with a hundred or so small areas. The computer has changed all that. Now it's a matter of less than one minute on the printer to turn out a map, 30 by 40 inches in size, complete with the data and shadings for 100 different areas. This kind of capability means that maps have moved out of the category of window-dressing and into the position of every-day working tools for spotting and interpreting the geography in the statistics.

In the urban statistical packages which we have designed for use by city planners and administrators, we routinely produce 100 or so different maps displaying selected tract-level data on population, housing, labor force, business activity and land use. And if anybody should want another 100 maps to be produced from other data in the file, it's only a matter of a few instructions to the computer.

A Shorthand Technique

Another kind of computer graphics is proving to be very useful in working with data for very small areas such as blocks. Here, we've developed a sort of shorthand technique whereby the computer prints an index number or a symbol at the position which corresponds to the geographic center of each of the blocks on a map. When you overlay this print-out with a transparency which shows the outlines of blocks, tracts or planning areas, you have a very effective way of looking at the geography of block data. It's also very fast — a standard high-speed printer can rattle off such a map, completely plotted with the data for 2,500 blocks, in a matter of 5 or 6 seconds.

This Model-T form of grid coordinate plotting is also an excellent device for spotting concentrations or exceptions in the data. All you do is select and print just those blocks which have values over a certain level. We used this approach in dealing with a good deal of our data for the 655-block West Side Detroit area which was struck by last summer's riots. In all, we printed over 100 maps showing concentrations of such items as total population by block, the number of housing units, the number of new movers, the number of female-headed households with children, etc. When we read the block printouts through overlays indicating those blocks where the greatest riot activity had taken place, it became pretty easy to spot some important relationships between data and damage.

Another new way to use computers for dealing with the geographic relationships in small area data is beginning to emerge. Somebody said to us not long ago: "Why don't you put all your maps on transparencies so that you could pile them on top of each other and get a cross-section look at what's going on?"

We thought about that one and reasoned that no matter how thin you printed them, you could only pile the maps so high. Then it occurred to us that the same computer programs that sorted out the tracts in order to set up the shadings on the maps could just as well assign a rating or rank number to each tract. Then if you had a rank number for each tract for a whole series of factors — such as the percentage of female headed household with children — you could add them up and get a composite rating by tract which would be the equivalent of trying to see through a bunch of transparencies.

One of the beauties of this approach is its tremendous flexibility. Rankings can be assigned, combined, regrouped and weighted with very little effort, and the composite ratings can in turn be summarized and displayed on maps. This is a good example of the kind of multidimensional processing that typifies the new and growing use of the computer as a working, analytical tool.

"Mathematical Models"

The last in this particular list of small-area computer capabilities is the new and fast-growing field which includes prediction, simulation and that all-encompassing term, "mathematical models". Our first-hand knowledge and experience in this field are so limited that we're going to pass up this one except for a single comment: while computer capability makes many of these models possible, it doesn't necessarily make them good. There's even one school of thought which says that every model builder should be required to demonstrate the assumptions and prejudices in his handiwork on a scratchpad and a desk calculator before he's even allowed to get near a computer.

To sum up, then, our first general observation is that there exists right today plenty of computer capability to deal very effectively and very imaginatively with small-area data.

Comparing Data From Many Sources

Our second major point is that these capabilities are adding whole new dimensions to the *practical* use of statistical data. Some kind of geographic identifier — be it individual address, or the designation of an area with its block, tract, or ZIP zone number — is turning out to be the common denominator with which great masses of data from many sources — up to this point unmanageable and unmergeable — can now be related and compared and summarized in a most useful and practical fashion.

Here are a few quick examples of how the computer is taking data from several sources and putting it together in ways such that one plus one equals a lot more than two.

Example: The Car Business

Our first example is from the car business. In one computer file, we have an address-by-address listing of all the new cars that are purchased, month-by-month. In a second file, we have a listing of all car owners. And thirdly, in our City Directory tapes, we have a detailed demographic profile of all the families involved.

Step 1 is to relate car purchase by make with car ownership by make — for an area such as a dealer trading zone or census tract. The rule of thumb is that car purchase closely follows car ownership — the odds are very good that the owner of Make A will buy another Make A, at an almost predictable rate.

As soon as the computer starts calculating purchase rates as a percentage of ownership, it's on its way to smoking out the geographic areas where sales attention is due — the beginnings of broad-scale, small-area exception analysis. Incidentally, this kind of analysis led to the discovery that there are more car buyers in just the non-metropolitan areas of the state of Ohio than there are in 85 of the traditional top 100 car markets across the country.

Then when you add in the economic and the demographic profiles of the families in each area, you have a powerful new index of the kinds of people who are in or out of the car market, what they are or are not buying, and at what rate. Facts of this kind up to now have been available only through samples, which for reasons of both expense and administration are usually so thin as to be of limited value in small-area analysis.

And finally, if you want to go way behind all the numbers and get into such things as opinion and motivation, you have in these merged and multi-dimensional computer files a well-structured and broadly based statistical universe from which you can make very precise selections for your sample.

Example: A Shopping Center

Let's take another example of new dimensions for statistics in space. Suppose that you are investing in a shopping center and that decisions on alternate sites are being made. The critical, money-making questions go something like this: How much business will the supermarket do, and how big should it be; how much floor space should we plan for a shoe store, and what grade of merchandise should it carry; and is there enough potential business to support a jewelry store? The problem is too much space and you lose money; not enough space and you lose business.

Studies are available which relate family expenditures for hundreds of different products and services to family-type — number in the group, occupation of the head, whether they own or rent their home, whether or not they have children, etc. Step 1 is to put the computer to work assigning and aggregating dollar expenditures by product, by family and by small-area — even down to the block. Next, set up the configuration and the weighting factors for the trading zones which surround each of the possible sites, and use the computer to make the thousands — or maybe even millions of calculations — that are required to estimate potential dollar sales by product line.

Finally, boil the whole thing down into as few as possible decision-type numbers and you have the exciting new pattern of geographic analysis that's beginning to emerge in practice as a result of computer capability in dealing with small-area data.

Example: Housing

A final example, this time from the public sector, is the critical problem of housing. Routine step one is to take an inventory of existing housing in the area involved — how much, what kind, what condition and where. Routine step two is to take a count and prepare a profile of the people involved — type and size of family, economic condition, etc. Then, with step three, we get into that new element of geography and computer capability. How does the inventory of housing and people — by location — relate to the needs for housing, present and future — by type and by location?

Where are — and where will be — the smokestacks and the jobs versus where are the workers? What is — and what will be — the journey to work, and the need for transportation facilities? How about schools, bussing, integration?

By putting together computers, small area data, and some proven, accepted analytical techniques and models, it's possible to come up with practical answers to questions like these — answers that would never have been possible just a few short years ago.

A Difficult Taskmaster

Our third and last point, you may remember, is that the computer is turning out to be a stern and difficult taskmaster for those who are pioneering in the use of small-area data.

One of the biggest jobs, we find, is defining the job. What are the questions that need to be answered with small-area data? What data are needed to get answers that are useful and not just interesting? What is the best way to process and summarize the information? Who is going to take on the job of doing the analysis and drawing the conclusions?

Over the last three or four years, we've seen some rather negative reactions to the availability of small-area data — ranging all the way from complete disinterest, to skepticism, to informed detachment. The kinds of facts we're talking about are something new, and to some potential users, they're disturbing — resulting almost literally in an attitude of "Let's not rock the boat by confusing ideas with the facts."

It's heartening to see that recently quite a different climate is developing. Without question, there's a rapidly growing awareness of the need and opportunity for the use of small-area data. Couple this with the computer and technical capability that already exists, and you have the makings of a quantum leap forward in our skills in dealing with urban data.

At the same time, we must remain aware that there are hazards in moving too fast, and especially in accepting computer output at face value simply because all the answers come out in neat columns, complete to 8 or 10 decimal points. As someone recently put it, "G-I-G-O used to stand for 'Garbage In — Garbage out,' but nowadays, it's starting to mean 'Garbage In — Gospel Out'!"

The Need for Vigilance

We must also be alert to the fact that the computer is an avid and voracious collector of information. It is an indiscriminate eclectic, that makes it very easy and tempting to add just one more item to the file on the chance that somebody, someday might want it. Unless there's another somebody who is continually vigilant in restricting the file to what's truly useful — not merely interesting — you run the very real risk that your data bank will turn into a prohibitively expensive and unwieldy data dump.

And finally, in this kind of climate there's a need for constant vigilance — amounting almost to suspicion — in reviewing and appraising the computer's output, particularly when it is performing forecasting or quasi-analytical operations. It's well to remind ourselves that the output's only as good as the raw data and the instructions that make up the input, and that sometimes the question is not so much the computer's capabilities as the capabilities of the people who are telling it what to do.

Somewhere near center, there's the good solid position of careful appraisal mixed with well-founded optimism in applying today's computer capabilities to today's data requirements.

Without question, today's capabilities for handling small-area data with computers are adequate and are pointed in the right direction. Our skills and speed are picking up. And tomorrow's potentials for powerful, low-cost and innovative uses of the computer are becoming increasingly more varied, challenging, and exciting. ■