188
2/28/79

DR. 2278

MASTER UCRL-50025-78-3

ELECTRONICS ENGINEERING DEPARTMENT QUARTERLY REPORT NO. 3—1978

Inside: Delay generator for hydrodiagnostics.
Framing-camera tube for sub-100-ps range.
Automatic analysis system used in NURE program.
Mass storage subnetwork with 1-trillion-bit capacity.
Exploratory data analysis of photochemical oxidants in
Imperial Valley, California

October 30, 1978

Work performed under the auspices of the U.S. Department of
Energy by the UCLLL under contract number W-7405-ENG-48.

**LAWRENCE
LIVERMORE
LABORATORY**
University of California / Livermore

LAWRENCE LIVERMORE LABORATORY
University of California Livermore, California 94550

UCRL-50025-78-3

ELECTRONICS ENGINEERING DEPARTMENT QUARTERLY REPORT NO. 3—1978

MS. date: October 30, 1978

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, for any of their contractors, subcontractors or their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

FOREWORD

The EE Department Quarterly Report is published with two purposes in mind: (1) to inform readers of various activities within the Department and (2) to promote the exchange of ideas.

The articles, by design, are brief summaries of EE work. For further details on a subject covered, please contact the individual listed at the beginning of the article in question: that person is primarily responsible for the content of the article. Inasmuch as most projects are the result of the cooperative efforts of many individuals, the article contact may either provide the requested information directly or refer you to the appropriate person to answer your question.

EE Department personnel are encouraged to submit articles for consideration to the Publications Committee. Committee members include:

R. A. Condouris--Technical Editor
T. Holdsworth--Field Test Systems Division
D. R. Dunn--Engineering Research Division
L. L. Reginato--Fusion Energy Systems Division
J. W. Spencer--Operations Division
W. F. Thompson--EE Department Staff
A. M. Kray--Nuclear Energy Systems Division
S. D. Winter--Laser Engineering Division

CONTENTS

New Delay Generator Combines Accuracy with Remote Programmability and Self-Testing	1
Framing-Camera Tube Developed for Sub-100-ps Range	5
Automatic Analysis System Analyzes Geological Samples for LLL NURE Program	10
New Mass Storage Subnetwork Provides On-Line Capacity of One Trillion Bits	15
An Exploratory Data Analysis of Photochemical Oxidants in the Imperial Valley, California	25

NEW DELAY GENERATOR COMBINES ACCURACY WITH REMOTE PROGRAMMABILITY AND SELF-TESTING

We have developed the DG-8 programmable digital delay generator to improve our computer-controlled hydrodiagnostic capability. Its eight output channels generate time delays from 100 ns to 10 ms in steps of 10 ns. Programming the DG-8 is done either from its front panel or from a remote terminal. Modular construction and a unique set of self-test routines make the DG-8 easy to assemble and maintain. This combination of accurate timing, remote programmability, and self-test has made the DG-8 an important contribution to our computer-controlled hydrodiagnostic work.

Delay generators play a central role in our hydrodiagnostic work. Every test requires that a number of events be triggered at set time intervals. To implement this, we generate a number of separate pulses at specific times after a single reference pulse occurs (Fig. 1). In the DG-8, a single, input reference pulse triggers the circuits of all eight output channels. Then each channel circuit counts out its own delay time and generates an output pulse. The accuracy of these time delays better all of our present and anticipated test requirements.

Simplified Operation and Maintenance

The DG-8 incorporates several features to simplify its operation and maintenance. We designed the front panel (Fig. 2) so that a new operator can set up the instrument without needing to refer to the instruction manual. Seven-segment numeric indicators display the delay-time settings of all eight channels. Other lights on the front panel prompt the operator on how to change these settings through the 12-key keyboard. In a computer-controlled test stand, these settings can also be changed from a remote terminal through the general purpose interface bus (GPIB).

Use of modular construction techniques eases instrument maintenance and self-test diagnostics. Incorporating a microprocessor facilitates the operation of the front panel, the GPIB, and the self-test routines.

Advanced Clocking Circuitry Assures Accurate Timing

Because timing accuracy is a vital concern in hydrodiagnostics, the DG-8 incorporates advanced

clocking circuitry (Fig. 3). After receiving the input trigger, a specific number of clock pulses are counted and then the output pulse is generated. By using a 200-MHz oscillator for the clock, the time from the input trigger pulse to the first clock pulse is repeatable to within 5 ns. Variations in the time delays of the circuit components are manually "zeroed out" with a 0- to 30-ns adjustable delay line.

The accuracy of the generated delay time depends on the accuracy of the 200-MHz clock. Because this clock is phase-locked to a 10-MHz reference, an external 10-MHz frequency standard can be used to generate extremely accurate delay times. In a bunker where we conduct our hydrodiagnostic tests, several instruments are clocked by this external standard to give very accurate timing throughout the hydrodiagnostic system.

Microprocessor Central In Programming, Control, and Self-Test

The DG-8 incorporates a microprocessor that makes the instrument easy to operate and maintain. Also, using the microprocessor minimizes the amount of hardware and makes future enhancements relatively easy to implement.

The microprocessor has three functions:

- Handle the front panel keyboard and numeric displays.
  - Process all GPIB operations.
  - Control the instrument self-testing (Fig. 4).
- The 12-key keyboard (0-9, decimal point, ENTER) allows the operator to program the instrument from the front panel. The delay time for a channel is set by first entering the channel number (1 to 8) and then entering the delay time (0.1 to 9999.99) in microseconds. Pushing a special combination of keys instructs the microprocessor to test its delay generation circuits. The 48 seven-segment numeric displays on the front panel display the delay time of

Fig. 1. The delay generator is used to trigger a number of events at set time intervals. One DG-8 can trigger up to eight such events.

Fig. 2. The front panel allows a new operator to make all settings without referring to an instruction manual.

each of the eight channels. All of these displays are multiplexed, which means that only one set of segment drivers is needed. In a nonmultiplexed design, 48 sets of drivers would be required.

The microprocessor tests the delay generation circuitry in two different ways. Each channel has a status bit that goes low when the input trigger is received and goes high when the output pulse is generated. This bit should remain low for the amount of delay time that the channel is set to. The microprocessor measures the actual amount of time that the bit is low by reading the status of this bit continuously and by counting the number of its own clock cycles. If the delay-time setting and the delay-time

measurement do not correspond, then a circuit problem exists in that channel.

Another test involves setting all channels to the same delay time and continuously reading the status bits. If all bits change at the same time, then all channels probably work properly. Other test routines written in the future will be installed by simply replacing the erasable programmable read-only memory (EPROM).

Although the microprocessor plays a minor role in the generation of the delayed pulses, it plays the central role in programming and displaying the delay time and in testing the delay generation circuits.

Fig. 3. Advanced clocking circuitry assures accurate delay-time generation.

General Purpose Interface Bus Handles Data and Protocol

The general purpose interface bus (GPIB) deals with the transfer of data among instruments on a common bus. It defines the protocol that instruments on this bus must follow. The microprocessor is programmed to follow this protocol in manipulating the 16 bus lines. The microprocessor receives instructions to enable or disable the front panel keyboard and to set the delay time of a channel. It also receives instructions to send data. These data involve the keyboard enable, the delay time of a channel, and several status bits. Using the microprocessor makes the circuit design and implementation of this GPIB function relatively simple and easy.

Construction Emphasizes Reliability, Maintainability, and Low Costs

The assembly of the DG-8 emphasizes high reliability, high maintainability, and low cost (Fig. 5). All circuits are constructed on printed circuit boards. Except for the front panel, all printed circuit boards plug into connectors mounted on a mother board. All of the readouts and switches are soldered to the front-panel printed circuit board. Two flat ribbon cables carry all signals between the front panel and the mother board. This makes the DG-8 easy to assemble and maintain.

Because the output boards handle a large current change in a short amount of time, they are encased in a metal box to guard against electro-

Fig. 4. A microprocessor interfaces the front panel and GPIB to the self-test routines.

Fig. 5. The DG-8 is designed for high reliability, high maintainability, and low cost.

magnetic interference. These high voltage pulses travel from this box to the back panel in a coaxial cable. Also, to ensure high reliability all circuit components operate at less than one-fourth their maximum power ratings.

Applications

Table 1 lists the specifications of the DG-8 programmable digital delay. It can be used wherever delay times from 100 ns to 10 ms need to be generated. Because of our hydrodiagnostic system requirements, the DG-8 generates a 0- to 80-v output pulse. However with a minor modification, it can be made to generate any type of output signal. The only change is to replace the small output board with a circuit that converts the 0- to 5-V pulse from the counters to the desired output signal.

The advantages of the DG-8 over most other delay generators are its timing accuracy, its remote programmability, and its self-test capability.

Table 1. The DG-8 contains eight time-delay channels, a single trigger input, and eight independent outputs.

Input trigger:	15 to 150 V, 2000 Ω , 10 pF, dc coupled.
10-MHz input:	5 V, 3000 Ω , dc coupled.
Output pulse:	80 V, 50 Ω , 500-ns width at 40 V.
Internal 10-MHz oscillator accuracy:	$\pm 0.005\%$.
Time range, delay:	100 ns to 10 ms, programmable in 10-ns intervals.
Time delay jitter:	5 ns.
Mechanical dimensions:	19 by 5 1/4 x 14 in.
Power requirements:	105 to 125 V ac, 60 Hz, single phase, 1 A.

FRAMING-CAMERA TUBE DEVELOPED FOR SUB-100-ps RANGE

A new framing-camera tube, developed by Electronics Engineering, is capable of recording two-dimensional image frames with high spatial resolution in the sub-100-ps range. Framing is performed by streaking a two-dimensional electron image across narrow slits; the resulting electron-line images from the slits are restored into a framed image by a restorer deflector operating synchronously with the dissector deflector. We have demonstrated its performance in a prototype tube by recording 125-ps-duration framed images of 2.5-mm patterns. The limitation in the framing speed is in the external electronic drivers for the deflectors and not in the tube design characteristics. Shorter frame durations (below 100 ps) can be obtained by use of faster deflection drivers.

Comparisons of Streak and Framing Cameras

A camera system sensitive to plasma radiation provides more useful information from a plasma experiment than any other diagnostic technique. Consequently, high-speed photography techniques have been extensively used and refined. The principal tools that have evolved are the streak camera and the framing camera.

A streak camera makes a continuous time-record of a single slit-defined line in a particular event (Fig. 1). The slit aperture is kept open, and the image is swept onto a phosphor screen, which holds the image for recording on film. Different parts of the film record different times in the event being photographed; what results is a continuous time record in only one spatial dimension. This one-dimensionality is a disadvantage of the streak camera record. A particular line of the object can be viewed throughout its time history, but a two-dimensional spatial view of the object cannot be recorded.

A framing camera, on the other hand, records on each frame only the event occurring at a particular time (Fig. 2). It records a series of framed images, each two-dimensional in space; a strip of motion picture film is a series of framed images. The series progresses in time, but is not continuous; a time gap exists between each frame. In the past, commercially available framing cameras have not operated fast enough to record events of very short duration, such as laser target implosions. These cameras degrade the sharpness of the picture because the two-dimensional object moves within the recorded frame period.

This limitation has been encountered in LLL's laser fusion program, specifically in the measurement and characterization of laser-generated

plasmas. In a laser-implosion experiment, a target about 100 μm in diameter is imploded to a fraction of its initial size in about 100 ps. Both the spatial and temporal properties of the plasma resulting from the nuclear burn must be characterized, as must the laser beam. The most direct method of recording these experiments is with a camera system, one that provides the required space-time resolutions for imaging the implosion processes.

The simplified diagram in Fig. 2 illustrates the operating principles of dissection and restoration of in a three-frame tube of our design.¹ Each dissection line in the framed image is well focused and has a temporal resolution similar to that in a streak-camera tube. The framed image is scanned linearly in time; i.e., a time delay occurs between the first and last lines of the frame. If necessary, computation techniques could be used with a multiple-slit dissector tube to obtain single-time frames.

Multiple-framed images are best achieved when the slit width is made narrower than the spatial resolution in the framing-camera tube (FCT) and the restorer deflector is swept in the opposite direction as the dissector deflector. The framed-image aspect ratio can be adjusted by varying the sweep-voltage rate of the restorer deflector with respect to the dissector deflector. Proper delay between deflector drives must be provided for the electron transit time between the two deflectors.

Prototype Tube

We tested the concept of our FCT in an initial prototype using a thermionic cathode with a 25- μm -thick molybdenum shadow mask in contact with the cathode surface. The shadow-mask pattern, approximately 2.5-mm square, generated a similar electron pattern in the FCT. We use a 394-mesh/cm accelerating grid located 2 mm from the cathode to apply a high electric field (<3000 V/cm) at the cathode

For further information about this article, contact Ralph Kalihjian (Ext. 28618).

Fig. 1. Simplified drawing of a streak camera recording a laser-target implosion. The slit aperture is kept open throughout the entire experiment. The image, swept from bottom to top on the viewing screen, is one-dimensional in space. It records only a line image, that line defined by the width (shortest dimension) of the slit. The hourglass shape of the recorded image represents the implosion and subsequent expansion of the target.

Fig. 2. A simplified three-dimensional perspective of the FCT, illustrating the basic principles of dissection and restoration. (Numbers indicate electron line images.)

surface, thus minimizing the transit-time dispersion between the cathode and grid to less than 10 ps.

We have characterized the tube in two time ranges. First, we have used a conventional driver for the deflectors (giving an $0.5\text{-}\mu\text{s}$ frame period) to demonstrate the principles of line dissection of the electron image and subsequent restoration of the image frame by synchronized deflection. Second, we have used the fastest available driver in our laboratory (125-ps frame period) to give an indication of the FCT's capability in ultrafast framed-image recording.

The first tests were performed with sawtooth waveforms generated by commercially available oscilloscopes. The results of these tests showed that the tube can restore $0.5\text{-}\mu\text{s}$ framed images from electron images that have been dissected either parallel or perpendicular to the lines in the pattern. This demonstrates that high-resolution framed images can be restored independently of the direction of dissection. The results also showed that three $0.5\text{-}\mu\text{s}$ -period framed images could be dissected and restored from the three slits in the FCT. The FCT has a zoom lens that can vary the overall magnification from three to nine.

Faster deflector-voltage sweeps (faster rise times) giving 125-ps frame periods have been generated with a mercury-switch pulser. The $50\text{-}\Omega$ pulser was

used to drive three $125\text{-}\Omega$ lines in parallel for the accelerating grid and the two deflectors. Appropriate delays were introduced so that both deflection sweeps occurred within the gate pulse at the grid. To compensate for the electron transit time between the dissector and restorer deflectors, a 6.5-ns delay was also introduced into the transmission line used to drive the restorer deflector. The fastest-rise-time waveform available from the pulser (1400 V ns) was applied to the restorer deflector. We recorded our fastest picture — the 125-ps framed image (with a $3\times$ electron-optical magnification) shown in Fig. 3 — by using the mercury-switch pulser for a 2.5-mm pattern on the cathode.

In Fig. 3, the framed image can be compared with the shadow mask that generated the electron beam pattern. The framed image shows good spatial resolution, with a $67\text{-}\mu\text{m}$ -wide dark space between the right pair of lines. We recorded the framed image by using a three-stage, $250\,000$ -gain image intensifier (having unity magnification and 25-line-pair/mm spatial resolution) at a 60-Hz repetition frequency. We anticipate to record frame speeds in the sub- 100-ps range by using faster deflection drivers currently under development.

Two other prototype variations of our FCT have been fabricated for use in laser-fusion studies: a tube with an x-ray-sensitive gold photocathode

Fig. 3. Shadow mask used to generate the electron-beam pattern that is dissected and restored into a 125-ps framed image.

of 1-cm diameter on a 15- μ m-thick beryllium substrate, and a tube with an S-20 photocathode.

On the first tube, to avoid atmospheric rupture of the thin x-ray window, we assembled a vacuum-valve port integral to the tube. After the tube is mounted onto the fusion-plasma chamber (which is nominally at 10^{-5} Torr), the vacuum port is opened to allow direct x-ray irradiation of the photocathode. To dismount the tube from the plasma chamber, the valve port is closed before opening the chamber to atmospheric pressure. Figure 4 shows the x-ray FCT, without the vacuum valve.

The S-20 photocathode version of the FCT is very similar to the x-ray tube, except for the cathode assembly. A cut-away view of the tube is shown in Fig. 5. An avalanche transistor pulser² has been designed to drive the deflectors of both these FCT's for application in the laser-fusion studies.

Conclusions

Our FCT is based on line-by-line dissection and restoration of the image. For 2.5-mm-square images at the photocathode, our initial prototype tube has given framed images of 125-ps duration and a spatial resolution better than 65 μ m. The limitation in frame period is in the deflector drivers and not in the tube design.

Temporal resolution per resolvable image line is comparable to that in a streak-camera tube. The current tube can image a resolvable line every 2.7-ps. The photocathode versions of the FCT are currently being evaluation in laser-fusion studies.

Fig. 4. Prototype x-ray FCT (before mounting on vacuum station). Overall length of FCT is about 0.5 m.

Fig. 5. Prototype FCT with S-20 photocathode. The design, except for the cathode assembly, is the same as for the tube shown in Fig. 4.

where they will be used to record plasma images and laser-beam characteristics.

Further FCT development involves the design of a five-frame tube. Ultimate framing speed in the tube itself will depend on minimizing the transit-time dispersion in the tube and on adapting wide-bandwidth deflectors in the tube to reduce the deflection-driver power requirement.

References

1. R. Kalibjian, "100-ps Framing-Camera Tube," *Rev. Sci. Instrum.* **49**, 891 (1978).
2. S. W. Thomas, G. R. Tripp, and L. W. Coleman, "Ultrafast Streaking Camera for Picosecond Laser Diagnostics," in *Proc. 10th Inter. Congress High Speed Photography, Nice, France, 1972*.

AUTOMATIC ANALYSIS SYSTEM ANALYZES GEOLOGICAL SAMPLES FOR LLL NURE PROGRAM

We have developed a completely automated system to analyze geological samples for the National Uranium Resource Evaluation (NURE) program. The NURE program was established by the Department of Energy (DOE) to identify regions of promising uranium presence in the United States. Electronics Engineering designed and built an identification coding scheme, a transport control system, and neutron and gamma-ray data acquisition equipment to analyze the NURE samples.

As part of the NURE program, LLL is one of four DOE laboratories systematically studying uranium distribution in soil, water, and lake and stream sediments. We are specifically responsible for analyzing 200 000 earth and water samples from seven western states: Arizona, California, Idaho, Nevada, Oregon, Utah, and Washington.

The procedures for surveying and preparing samples have been standardized. Contract and LLL geologists gather the samples according to a grid pattern. The soil samples are dried, and a punched card is produced to identify the source of the sample.

The solid samples are analyzed by the instrumental neutron-activation analysis (INAA) system, and the water samples are analyzed by an emission spectrometer system, called inductively coupled argon plasma (ICP) source. Results from these analyses are processed in the Livermore timesharing computer system (LTSS) to produce reports for NURE headquarters, Grand Junction, Colorado.

INAA and ICP Instrumentation

We chose INAA and ICP because they are amenable to automation. The Livermore Pool-Type Reactor (LPTR) provides neutrons to activate or produce radioactive isotopes in the soil samples. However, we had to build a system to transport the samples to the reactor core area for irradiation and then to neutron and gamma-ray detectors for acquiring data for subsequent analysis in LTSS.

Samples that are to be neutron activated are placed in polyethylene capsules (called rabbits), which contain the samples while they pass through a microcomputer-controlled pneumatic transport system. Rabbits containing samples are loaded into a dispenser attached to the pneumatic system.

From this point on, the system becomes a fully automated hands-off system.

The ICP system for analyzing water samples comprises a commercial emission spectrometer and an LLL-developed microcomputer control and data acquisition system. The microcomputer stores and loads computer programs to a PDP-8 mini-computer in the emission spectrometer. The microcomputer then acquires data from the spectrometer system and records the data on floppy disks for subsequent analysis in LTSS. The remainder of this article concerns the automated INAA system and how it is used to analyze solid samples.

Microcomputers Control the Entire System

DEC LSI-11 microcomputers control all parts of the INAA system:

- Rabbit identification code driller and verifier.
- Rabbit identification code and punched card reader.
- Send station sample loader and identification code reader.
- Rabbit transport system.

The rabbit identification is a binary code drilled in the circumference of the sample capsule (Fig. 1). After the drilling operation, the binary code is verified by sensing a laser beam reflected by the rotating sample rabbit. We chose the drilled ID code method because it provides a system that can be automated and that does not add any contamination to the sample rabbits as would an ink-coded system.

A microcomputer controls a pneumatically driven drill motor and a stepper motor to drill 2 synchronizing holes and up to 21 identifying holes in a 1, 2, 4, 8, ... binary code. Up to 2 097 152 different ID codes can be generated by the 21-hole identifying scheme. Depending on the number of holes drilled, the microcomputer-controlled driller can drill and verify about 90 rabbits per hour.

For further information about this article, contact John Baker (Ext. 26597).

Fig. 1. Rabbit identification code reader. The drilled holes represent a binary-coded ID, and the photosensor translates the reflected light into binary light pulses.

After the rabbits are drilled, cleaned, and loaded with soil samples, they are taken to the rabbit ID code and punched card reader. Here, the rabbit ID code is read and correlated with sample information from punched cards prepared for each sample. The rabbit ID code data and sample information are stored on a floppy disk for subsequent entry into the ITSS.

Finally, the samples are ready for neutron activation and for neutron and gamma-ray analysis in the INAA system. They enter the INAA system through the send station. In the send station (Fig. 2), the rabbits are loaded from a magazine hopper, the rabbit ID code is read, and the rabbits are propelled by nitrogen gas to the reactor core area for neutron irradiation for a predetermined time. A microcomputer controls the sequencing and timing and detects error conditions in the transport system. Diverter switches are operated to route rabbits as in an automated railway system. Microswitches interfaced to the microcomputer verify the operation of the diverter switches. Photosensors verify the correct passage of rabbits to the reactor core, to delay stations, to the delayed-neutron detector, to the four gamma-ray detectors, and finally to rabbit storage or rabbit dump areas.

Rabbits are automatically sent to the lead-shielded storage area for a few days to permit safe storage of the samples while the induced radioactivity decays. After about 10 days, the gamma rays in the samples are counted either by reintro-

ducing them into the INAA system or by a separate gamma-ray detector system. This second analysis is required in order to analyze for long-lived isotopes that may have been masked by highly radioactive shorter-lived isotopes during the first gamma-ray counting. The results of the data analysis and the geographical data of the samples are then used to produce maps showing uranium concentration.

In case of a system malfunction, rabbits are automatically sent to the lead-shielded dump area. After the radioactivity of these samples decays, they are re-entered into the INAA system for neutron and gamma-ray analysis.

200 000 Samples by 1982

We optimized the microcomputer transport timing for a maximum throughput of 17 samples per hour. Analysis takes place 10 to 24 hours per day 5 days per week. Thus a total of 200 000 samples will be analyzed by 1982 when the program is expected to finish.

Software Development

Writing software for the microcomputers was one of the major tasks of the development phase. The software tasks were split into modules so that the software development could be shared by several electronics engineers. We used a high-level language, FORTRAN IV, as much as possible to speed the development and to make the computer programs easier to read and maintain.

Gamma-Ray Spectra and Delayed-Neutron Data

Microcomputer systems acquire delayed-neutron and gamma-ray data. Neutron data acquisition through neutron activation and delayed-neutron counting is the primary means of detecting uranium in the samples. However, gamma-ray detection also provides uranium concentration data and data on other elements that are often found near uranium deposits. We use the INAA gamma-ray analysis to determine the concentration of up to 35 elements in the soil samples.

The delayed-neutron data is collected by a calibrated BF_3 gas-filled detector and an event counter interfaced to the transport control system microcomputer.

Fig. 2. Layout of rabbit transport system.

To acquire the gamma-ray data, we developed and built four 4096-channel pulse height analyzers (PHAs). Each PHA is composed of a microcomputer and an interface to a commercial analog-to-digital converter (ADC). The PHA functions are generated by software in the microcomputer (Fig. 3). Gamma rays from the neutron-activated soil samples strike lithium-drifted germanium [Ge(Li)] semiconductor detectors. This produces current impulses that are amplified and digitized by the ADCs. Up to 10 000 gamma-ray events per second can be handled by the PHAs. They internally produce a spectra of events versus energy levels. These spectra are used to determine the isotopic makeup of each sample as each isotope produces a characteristic gamma-ray spectra. The four PHAs are interfaced to a single commercial graphics cathode ray tube (CRT) terminal for observing proper PHA operation. The software for each of the four PHA computers is loaded from a fifth microcomputer, which acts as a master control for the entire analysis system.

Central Microcomputer Ties System Together

The rabbit transport system and the four PHAs are tied together in a star network, with a master control LSI-11 microcomputer at its

center (Fig. 4). The master control microcomputer has a floppy disk unit for program storage. Master control can load operating programs to the four PHAs, send scheduling parameters to the rabbit transport microcomputer, receive status information from the rabbit transport and data from the PHAs, and record the data on disk memories. The communication hardware between the master control and PHAs can handle 40 000 bytes s (320 000 bits s). In actual operation, the communication rate is about 4000 bytes s, with comprehensive error checking to ensure that the data transfer is 100% accurate.

The microcomputer network is completely automated, requiring only the entry of scheduling information and gamma-ray detector geometry at startup time. From that point on, it can run unattended, printing out status reports, irradiating and analyzing samples, and storing data on the disk memories until it runs out of samples to be analyzed or a hardware malfunction occurs, which would cause an orderly automatic system shut-down.

Conclusions

The microcomputer-controlled INAA system has been very successful in analyzing more than

Fig. 3. Gamma-ray detector and pulse height analyzer system.

Fig. 4. The entire NURE INAA network is controlled by one LSI-11 microcomputer with a floppy disk storage system.

30 000 samples for uranium and 34 other elements. Some of the engineering decisions that proved favorable were:

- Use of a separate microcomputer for each major system function.
- Microcomputers connected in a star-type network.
- Programming in higher-level language, specifically FORTRAN IV.
- Drilled binary code for rabbit identification and a laser-photodetector for reading the code.

- Microcomputer-based pulse height analyzer.

In addition to the success of the system and the valuable experience that we gained, at least two technological spinoffs have resulted. The microcomputer communication interfaces that we developed are used extensively at the Nevada Test Site (NTS). Also, we are using a copy of the microcomputer-based PHA at NTS for analyzing gas samples from underground experiments.

NEW MASS STORAGE SUBNETWORK PROVIDES ON-LINE CAPACITY OF ONE TRILLION BITS

The multi-access storage subnetwork (MASS) is the latest addition to the Octopus computer network at the Laboratory. The subnetwork interfaces a Control Data 38500 Mass Storage Facility to the Laboratory's large, worker computers. The new data storage facility provides an on-line capacity of 1 trillion bits of user data. The MASS architecture offers a very high performance approach to the management of large data storage, as well as a high degree of reliability needed for operation in the Laboratory's timesharing environment.

The multi-access storage subnetwork (MASS), designed and developed jointly by Electronics Engineering's Computer Center Group and Computation Network Systems Division, combines state-of-the-art digital hardware with an innovative system philosophy. The project represents an investment of more than \$2.5 million in equipment and engineering labor.

The key LLL design features of the subnetwork that contribute to the high performance include:

- A data transmission scheme that provides a 40-Mbyte/s channel over distances of up to 1000 ft.
- A large metal-oxide-semiconductor (MOS) memory buffer controlled by a 24-port memory multiplexer with an aggregate data rate of 280 Mbytes/s.
- A set of high-speed microprocessor-based controllers driving the commercial mass storage units.

Reliability of the system is provided by a completely redundant network, including two control minicomputer systems. Also enhancing reliability is error detection and correction in the MOS memory. A hardware-generated checksum is carried with each file throughout the entire network, insuring integrity of user files. (A checksum is a sum used in an error-detection check before and after each transaction to verify that no digits changed.)

Background of Mass Storage at LLL

The Laboratory computer center has been using mass storage for several years, with on-line storage available to Octopus users.¹ (Octopus is the computer network for the Livermore timesharing system.²) To improve the service of this very useful resource, the Computation department initiated the acquisition of a new storage facility, as well as

developing a new subnetwork for implementation into Octopus.

The current mass storage device is the IBM Photo-Digital Store,³ which is connected to the worker computers through the file transport subnetwork.⁴ (Fig. 1). The Photo-Digital Store (called Photostore at LLL) was developed in the early 1960s and has been in operation at LLL nearly 10 years. This photographic data recording process has given our computer center a very reliable storage media that is extensively used. The Photostore has an on-line data capacity of 1×10^{12} bits. Seven trillion (7×10^{12}) bits of data have been recorded, with the seldom-used and inactive files placed off-line on the "shelf." The files currently "pointed to" by the file directory structure contains about 3.5 trillion bits.

Even though this storage device is heavily used and the recording characteristics are truly archival, it presents some operational problems:

- Daily preventative maintenance schedules of 3 hours per day is usually needed.
- Unscheduled emergency maintenance decreases effective availability to users.
- The mechanical, pneumatic, chemical nature and age of the device account for an increasing amount of down time.
- Limited throughput of data files because of the access mechanisms and read/write channels sometimes cause a bottleneck. Clearly, in a time-sharing environment, rapid accessibility of mass storage store files is paramount.
- Lack of modularity of Photostore requires the entire storage device to be taken off-line for emergency and preventative maintenance.

These inherent problems of Photostore, coupled with the inevitable retirement due to old age, has forced us to acquire a new storage device.

Apart from the Photostore, another drawback of the present system is the limitation of the architecture of the file transport subnetwork. The effective data rates of the subnetwork are severely constrained due to the interaction between system

---

For further information about this article, contact Donaldson L. Wentz (Ext. 24263).

Fig. 1. Current LLL file transport system.

activity on the 7600 and the transfer of data from the 7600 disk system to the PDP-10 memory where data is staged before being written onto the Photostore. The memory size limitation of the PDP-10 also impacts the effectiveness of the subnetwork.

For these reasons, we decided to engineer a new subnetwork to support a new storage device that addressed the problems of the old system.

CDC 38500 Mass Storage Facility

With the need to replace the Photostore, we wrote a set of specifications to acquire a new, commercial mass storage device. The specifications reflected the features that we felt were most desirable for the LLL computing environment: modularity, high media throughput, multipath access for availability, and a reasonable high aggregate data rate.

Control Data Corporation was the successful bidder with their 38500 Mass Storage Facility (MSF).

The CDC 38500 MSF is a high-density magnetic-tape library system. The magnetic tape, contained in a small plastic cartridge, is 2.7 in. wide and 150 in. long. Up to 2000 cartridges are stored in a cartridge storage unit (CSU). We purchased eight CSUs for a total cartridge capacity of 16 000. The CSU (Fig. 2) is an X-Y matrix with a mechanical pneumatic selector arm that selects and delivers a cartridge to one of two mass storage transports (MST) per CSU (16 MSTs are being provided).

The MST (Fig. 3) is the tape transport where the magnetic tape is pulled from the cartridge, into a vacuum column and around a capstan drive. A movable recording head records nine tracks of data in eight stream pairs, for a total of 144 tracks (Fig. 4). The tape is "shoe-shined" over the capstan with eight streams recorded in one direction and

Fig. 2. Cartridge storage unit (CSU).

eight streams recorded in the other direction on about 100 inches of tape. The recording technique is a slight modification of the ANSI standard group coded recording (GCR) with a density of 6250 bits per in. At this density, the nominal capacity of one cartridge is about 64 million bits. This makes the capacity of all eight CSUs at 1 trillion (1×10^{12}) bits. The same number of bit storage as Photostore.

The basic control for the CSUs and MSTs is a microprocessor-based controller called a mass storage adapter (MSA). Our engineering interface began with this MSA.

Figure 5 shows the CDC 38500 mass storage facility as purchased by LLL. Table 1 summarizes the basic characteristics of the MSF, and Table 2 compares the Photostore and the MSF.

Interconnection Subnetwork

Once a mass storage device was specified and selected for use in the Octopus network, we had to provide a means for interconnecting the device and

the major worker computers (Fig. 6). The multi-access storage subnetwork (MASS) was developed for this purpose.

The planning for MASS addressed the shortcomings and problems encountered with the older file-transport mass storage link. We configured the new structure so as to permit all of the storage elements to be accessible by all the major computers, and that use of the system provides a minimum impact on the operational load of the computer systems. MASS also was designed with reliability and expandability as key factors.

CDC 7600 and MASS Interconnection

Figure 7 shows in simplified form the interconnection between the CDC 7600 and MASS. The diagram illustrates the flow of data between the 7600 and the CDC mass storage facility. The major drawback in the existing file transport connection to the 7600 (Fig. 1) is that the data has to be allocated in small blocks and transferred from the

Fig. 3. MASS storage transport (MST).

819 disks, through the 7600 central processing unit (CPU), to the file transport peripheral processing unit (PPU). With MASS, we designed a high-speed communications channel that permits data to be taken off the 819 disks and shipped directly to MASS, bypassing the 7600 CPU. With sufficient memory allocation in MASS, a large block of data can be removed from the disk without the worry of head repositioning.

The communication channel between the 7600 and MASS consists of two LLL-designed devices. At the 7600 end is the worker data channel (WDC), which interfaces directly to the two PPUs that handle the 819 disks. At the MASS end, is the MASS data channel (MDC), which ties into the MASS data buffer and the controlling minicomputers (a set of Texas Instrument 980As). The two devices are connected together by two identical communication logic boards.

The communications channel operates at greater than 42 Mbits/s over distances of up to

1000 ft. This is necessary because the 38500 MSF, as well as MASS, are housed in a building different from where the four 7600s currently are. The communicating channel itself consists of 14 twin-lead, low-loss, double-shielded coaxial cables. Data is transferred in a burst mode at 42 Mbits/s, half duplex over 9 bidirectional lines; 8 for data and 1 for parity. The data bandwidth of the channel is sufficient to handle the maximum transfer rate of the 819 disks. First-in, first-out (FIFO) storage is used in both the MDC and WDC for smoothing the data flow between devices. The remaining lines of the channel are essentially unidirectional and are used for strobe, control, and status.

Once data is transferred over the communication link, it is buffered in the MOS memory provided in MASS. The intermediate buffering is necessary to compensate for the differences in the data transfer rate between the 7600 disks and the 38500 MSF. Since the memory has a 450-ns cycle and is interleaved, a maximum bandwidth of 280 Mbits/s into and out of the memory is possible. An LLL-designed memory multiplexer permits up to 24 data ports into the memory.

We designed the multiplexer to match the bandwidth of the memory. Its input data ports are capable of handling the 40-Mbits/s communication traffic, providing the total aggregate bandwidth of all the channels does not exceed 280 Mbits/s. Data is received, 16 bits parallel, plus 2 bits parity at one of the 24 multiplexer channels. The multiplexer and the memory control strip off the parity bits and form a 64-bit-wide word from four of the channel words. It then stores the 64-bit word along with an 8-bit Hamming code into the memory. Upon reading the data word, the multiplexer uses a Hamming code for single-bit error correction and double-bit error detection. During the read, the 64-bit word is reduced to four 16-bit words, and the two parity bits per 16-bit word are regenerated.

After immediate buffering in the MASS MOS memory, user data can either be transmitted to a CDC mass storage device or transported to another worker computer. If the data is to be transmitted to a 38500 MSF, the transfer will be handled by an LLL-designed MSF interface controller (MIC). The mass storage adapter (MSA) of the 38500 MSF was designed to attach to an IBM-type disk controller, which normally resides in an IBM 370 environment. The MIC replaces such a controller and permits easy access to the 38500 MSF by MASS. The heart of MIC is a high-speed Schottky bipolar microprocessor, the Intel 3000 bit-slice system. The basic cycle time of the 8-bit microprocessor is 200 ns. The microprocessor handles all the

Fig. 4. Tape is formatted in 16 streams; 8 streams recorded in each direction. Each stream is 9 tracks wide. Moving-head assembly covers two stream positions. Stream is broken into data blocks separated by an interblock gap. Data block is from 1 byte to 60K bytes in length.

command and status protocols of the MSF channel, as well as the actual data transfers. In addition to the normal read or write activity, the MIC also has a data-verify mode, which will compare data read off the 38500 MSF tape with data read from MASS MOS memory.

A pair of TI 980A minicomputers control MASS; one minicomputer is the master control, the other for backup and diagnostic activity. Each minicomputer has its own fixed-head disk system for 38500 MSF operational parameter storage, 38500 MSF statistics, and program storage. Associated

with the TI 980A, is an LLL-designed input, output (I/O) and direct-memory-access (DMA) expansion of the TI 980A bus system. Control information will be passed over the I/O bus, while data and system message traffic will be exchanged over the DMA.

To permit computer access to the MASS memory buffers, we designed a microprocessor-controlled block transfer device (BLT). This device permits message exchanges between the TI 980A and the worker computers over the worker communications channels, as well as data exchanges between the 38500 MSF and TI 980A for data verification.

MST — MASS storage transport

Fig. 5. CDC 38500 mass storage facility (MSF). Total capacity is 1 trillion data bits.

The BLT also proves to be a very valuable diagnostic tool. The T1 980A minicomputers are connected to the PDP-10 system via asynchronous data links through the Octopus message subnetwork.

Polynomial-Checksum Generators

An important feature of the MASS structure is the hardware polynomial-checksum generators and checkers built into the devices of the system. A 64-bit checksum will be appended to all 38500 MSF data blocks at the WDC and carried through the system. Since it is appended to user data, it is treated as data in the MASS buffer memories and stored on tape as actual data. The checksum will

be checked by both the MDC and MIC when the data passes through these devices. The checksum is a 64-bit polynomial that will cover a maximum data file of over 99 million bits.

MASS-Octopus Connection

Whereas the 7600 workers will have duplicate, direct paths to MASS, they will be somewhat different for the Cray 1 and the CDC Star 100s. Computations is currently installing a new commercial networking scheme built by Network Systems Corp. (NSC). It is called Octopus by LLL and will interconnect the workers to the associated resources. NSC also makes an adapter that connects

the 7600 PPU to the Octobus. We are currently modifying the adapter to interconnect a WDC-to-MDC link for communications with MASS (Fig. 8).

MASS was designed as a completely dual system, providing full redundancy (Fig. 5). Each worker computer has two communication channels to MASS. Both MASS memory buffers are capable of simultaneous operation, and in the event a memory element is lost, the operational memory unit will be able to handle the load, which will be determined by the master control TI 980A. This may result in some degradation of service, especially during peak activity periods, but the mass storage will still be accessible by the worker computers. Also,

the TI 980A I/O bus is serviced through a software-controlled reconfiguration panel, which will permit either an operator or the TI software operating system to reconfigure the MASS devices to either one of the two TI 980A I/O buses.

Control Philosophy

The job control system for communications between MASS and the worker computers will reside in the PDP-10 computers that currently support the file transport subnetwork. The job control system will consist of two software modules. One will be a communications module, which will handle all message traffic among the worker computers, the PDP-10, and MASS, and will schedule all data transfers. The other will be the file directory system, which will contain all the pointers for the user files recorded on the 38500 MSF.

When a request is made to make a transfer to or from MASS, the communications code would take the request from the worker computer and transfer the information to the MASS computer after fetching the necessary pointers from the file directory. The MASS software operating system on the TI 980A would then allocate the necessary space on a cartridge within the 38500 MSF and take action to mount that cartridge. Meanwhile, the logical cartridge number, along with stream and block numbers, would be transferred to the file directory on the PDP-10 computer, if the access involved a new pointer. The TI 980 software operating system would then transmit a message to the appropriate worker computer over its MASS communications channel to enter a request in that 7600's disk job

Table 1. Characteristics of CDC 38500 MSF.

Storage capacity:
16 streams/cartridge
4 Mbits/stream
64 Mbits/cartridge
2000 cartridges/storage module
1.28×10^{11} bits/storage module
Maximum transfer rate:
7 Mbits/s
Worst-case cartridge access:
11.0 s
Recording technology:
Magnetic tape
Nonreturn-to-zero inverted (NRZI) encoding
Bit density = 6250 bits/in. (9042 flux reversals per in.)
Track density = 50 tracks/in. (144 tracks)
Cartridge recording area = 100 by 2.75 in.
Tape speed = 129 in./s
8-position read/write assembly (20-ms maximum seek)

Table 2. Comparison of IBM Photostore and CDC 38500 MSF.

	Photostore	CDC 38500 MSF
Storage capacity	10^{12} bits	10^{12} bits
Recording technology	Photographic (nonerasable)	Magnetic
Unit of storage media	Cell (140×10^6 bits)	Cartridge (64×10^6 bits)
Device controllers	1	4
Readers	2	16
Recorders	1	16
Recording rate	0.25 MHz	3.0 MHz
Reading rate	1.50 MHz	4.9 MHz
Aggregate data rate	1.75 MHz	12.0 - 20.0 MHz
Independent X-Y selectors	1	8
Average access time/selector	4 s	9 s
Average access time (distributed over all selectors)	4 s	1.13 s

Fig. 6. Interconnection of worker computers to the CDC 38500 MSF.

queue. The data would then be transferred to, or from, the MASS MOS memory. Although a user's file can be as long as 64K million bits, the cartridges will be blocked into 60K-byte blocks, each block containing several header bytes and the 64-bit checksum. The MASS MOS memory will be dynamically allocated — depending on file size and MASS activity.

Because the CDC 38500 MSF tape is a rewriteable media, and considering that "shelf" storage for the CDC cartridges would be very expensive, a different approach to space allocation must be taken from that used with the Photostore. In the Photostore, user storage was essentially unlimited. The approach most likely to be used with MASS is to provide a group of users (application program, divisions, departments, a group, or an LLL project) a guaranteed amount of storage space and to provide sufficient tools to allow such a user indirect management of that space. Users will not be granted the privilege of knowing what physical cartridges belong to them, or in what CSU their files are stored. They will only be guaranteed a certain amount of "logical" cartridges or portions of cartridges. The purpose of this restriction is to guarantee maximum efficiency

and security in the management of the MASS file space.

Conclusions

MASS provides the Octopus network with a sixth major subnetwork with very reliable on-line data storage capability of up to 1 trillion bits.

MASS and the associated CDC 38500 MSF is in final software checkout. It is expected to be in full production in early 1979.

A MASS-to-Octopus link is currently being debugged, and software and firmware is being developed to provide both the STARs and Cray-1 access to the mass storage, as well as any future worker computer connected to Octopus.

In the future, MASS will be just one resource in a hierarchy of storage devices available to the Laboratory user. We foresee that the CDC 38500 MSF will be a medium-term storage device supplemented by a high-capacity disk farm for a very short term, high-activity files, and a truly archival storage device available for more permanent records.

Fig. 7. Partial diagram of the LLI. MASS interconnection between a 7600 and the CDC 38500 MSF.

Fig. 8. Multi-access storage subnetwork (MASS) (shown within dotted line).

AN EXPLORATORY DATA ANALYSIS OF PHOTOCHEMICAL OXIDANTS IN THE IMPERIAL VALLEY, CALIFORNIA

The Department of Energy (DOE) has designated I.I.I. as the lead laboratory in a comprehensive program to assess the impact of geothermal energy in the Imperial Valley, California. The Electronics Engineering department in support of the Laboratory's Division of Environmental Sciences has contributed to this Imperial Valley Environmental Program (IVEP), both in data collection and data analysis. Instrumentation for six fixed-site air quality monitoring stations and a mobile van were designed, developed, deployed, and maintained by electronics engineers and technicians. Part of our mandate for the program is to summarize the some 1.5 million measurements in a form suitable for dissemination to industry and government. To do this, we applied a number of interesting graphical and statistical techniques, which may be applicable to other programs.

The analysis of data can be divided into two broad categories: exploratory and confirmatory. *Exploratory analysis* is used to extract and explicate the informational content of some body of data, to hunt for patterns, and to summarize appearances. *Confirmatory analysis* is concerned with the acceptance or rejection of some hypothesis and an assessment of how strong the evidence is for that conclusion. The distinction between exploratory and confirmatory analysis is similar to the difference between police detective work and the formal judicial process, respectively.

This article reviews some of the data analytic tools used to conduct an exploratory analysis of air quality measurements from the Imperial Valley, California. We concentrate mainly on the analysis of photochemical oxidants because these data have turned out to be the most interesting.

These analytic tools are not necessarily limited to analyzing air quality data. They may also be used in other fields, such as economics, biology, sociology, and high-energy physics. We do not cover all or even most of the tools available. We do, however, concentrate on two major themes that support these techniques: graphical displays and resistant estimation. The LLL graphics facilities, which can produce high-resolution hardcopy and 35mm film, are especially appropriate to the use of these analyses.

Instruments for Measuring Air Quality

Ambient air is drawn in at a height of 5 m through a glass and Teflon sample train to the air

quality instruments. The outputs from these instruments and from other meteorological sensors are fed into a digital data-logging system that features a dual, cassette tape recorder. All signals are averaged over a 15-min time period.

Ozone concentration in the air is measured by the method of ultraviolet absorption. Nitric oxide (NO) and nitrogen dioxide (NO₂) concentrations are measured by the chemiluminescence technique. An important characteristic of these methods is that an instrument failure usually results in a low reading and not a missed reading.

Selection of Air Quality Monitoring Sites

The Imperial Valley, located in Southern California adjacent to the Mexican border, is a flat desert, rural environment under extensive irrigation and agricultural cultivation. The largest population center contains only 3200 people.

Six sites in the valley were chosen for the location of air quality monitoring stations. We selected these sites with regard to the most likely geothermal resource points, the expected ambient air pollution levels, the distribution of population centers, and meteorological conditions. There were a total of six fixed-location monitoring stations.

Problems in the Analysis of Air Quality Data

Air quality measurements have a number of characteristics that complicates their analysis and can often lead to incorrect interpretations. Many of the usual statistical assumptions do not apply, and we are forced to abandon the use of standard statistical techniques.

For further information about this article, contact Michael C. Axelrod (Ext. 20929).

Air quality observations are not generally distributed according to the Gaussian formula. The log-normal distribution has been a popular model for many data sets, but, as we shall see, it is the square root of ozone concentration that tends to follow the Gaussian distribution rather than the log of concentration.

The true multivariate nature of air quality data sets is often ignored, because multivariate statistical techniques are not nearly as well developed as the univariate methods. The inability of humans to visualize surfaces in dimension higher than three is a great obstacle to discovering "law-like" relationships in multivariate data sets. The detection of multivariate outliers is a difficult problem, one for which there is no general solution. Outliers are readings that behave very differently than the average of the majority of observations. It is important to realize that even if all the marginal distributions of a multivariate distribution are all Gaussian, the whole distribution may not be multivariate Gaussian.

Another complication (sometimes ignored) is the fact that air quality data really consists of sequential samples from a time series. The samples are not independent, and the time series can not necessarily be assumed to be stationary. The presence of serial correlation among samples can cause estimates of variability to be incorrectly low.

Another bothersome complication that can arise is the existence of gaps in the data record. These gaps can be caused by instrument failure, downtime due to maintenance and calibration, power outages, or external damage. We have experienced all these troubles in the Imperial Valley instrumentation.

Missing values can cause false impressions when comparing the data from various sites. Suppose two sites have nearly the same ozone concentration distributions, but one site has record gaps during the summer months when readings tend to be high. The sample distributions would look different, and we might incorrectly conclude that ozone concentrations at one site tend to be higher than the other.

Another difficulty is the presence of "wild" values or outliers. The existence of these outliers seems to be inevitable and is common to all large air quality data sets. There are two sources of these outliers, those associated with instrumentation or recording errors and those which are perfectly good data but are merely extreme. Extreme values may be caused by unusual environmental conditions or a freak emission. In some cases, it may be a great mistake to "throw away" extreme values without

good justification because their study may offer great insight into the mechanisms under investigation. We must employ techniques that recognize the possible occurrence of very extreme values.

Techniques of Resistant Data Analysis

A procedure is said to be *resistant* if a large change in a small fraction of the data does not significantly change the result of the procedure. The arithmetic mean $\bar{x} = (1/n) \cdot \sum_{i=1}^n x_i$ is not resistant; the change of only a single sample value can distort the mean considerably. For example, when

$$\{x\} = \{1, 2, 4, 8, 10, 16\}, \\ \bar{x} = 6.83.$$

However, when

$$\{y\} = \{1, 2, 4, 8, 10, 116\}, \\ \bar{y} = 23.5.$$

We see that the arithmetic mean is not a good measure of central tendency when outliers are present in the data.

Resistant Estimates of Location

Suppose $x_1, x_2, \dots, x_n$ is a random sample from a distribution F . We will assume that the x_i are independent and the distribution F has a center of symmetry. This center value is known as the *location*. Our goal is to estimate efficiently the location of F from the samples x_i , even when the sample may be contaminated.

An important and useful resistant estimator of location is the *trimmed mean*. To compute a trimmed mean, we delete a proportion of the data from both the high end and the low end, then calculate the ordinary arithmetic mean of the remaining values. A family of trimmed means can be defined by

$$T(\alpha) = \frac{W \cdot x_{n+1} + x_{n+2} + \dots + x_{n-\alpha} + W \cdot x_{\alpha+1}}{n(1-2\alpha)}$$

where

$$a = [\alpha \cdot n] \text{ and } W = 1 + a - \alpha n.$$

The weight factors W at each end are necessary to make the sum of the coefficients in the numerator equal to $n(1-2\alpha)$.

The trimmed mean is not just "throwing away data." The observations are first sorted by magnitude before the high and low values are deleted. The

extreme values still contribute to the estimate, but their influence is attenuated.

Note:

when

$\alpha = 0$; $\bar{I}$ is the ordinary *arithmetic mean*.

$\alpha = 0.25$; $\bar{I}$ is known as the *mid mean*.

$\alpha = 0.50$; $\bar{I}$ is a *median*.

We shall not dwell on the statistical properties of trimmed means; the interested reader should consult Ref. 1. When F is the Gaussian distribution, the mean is the best estimate of location. For contaminated Gaussian distributions, the midmean is generally preferable to the mean, which is in turn preferable to the median. If F is a symmetric distribution with long tails, the mean is the worst estimate. There are estimates of location that are superior to the trimmed mean and perform very nearly as well as the mean when the distribution is Gaussian. We will not go further than the trimmed mean in this article. A major disadvantage of the trimmed mean is that good data can be trimmed as well as bad.

Graphical Displays

We are now in a position to combine computer graphics and resistant estimation techniques to create some useful analytic tools. These tools are used on the photochemical oxidant data from the Imperial Valley energy project. Graphical displays are perhaps the single most important tool in exploratory data analysis. Figures 1 through 20 are reproduced exactly from our computer graphic output.

Box Plots

Box plots are a graphical technique used for displaying batches of data. Each batch is described by a five-number summary. These five numbers are: the maximum value, upper quartile value, median, lower quartile value, and the minimum value. We now construct a picture that represents this five-number summary, as shown in Fig. 1a. This picture reveals the location, spread, and skewness of the batch. We can then compare several batches by a simultaneous display of box symbols.

Fig. 1. Box plots of data. (a) Basic box plot of a batch of data. (b) Variable-width notched box plot. The width is proportional to the square root of the size of the batch.

The basic box plot can be subject to misinterpretation because the viewer may draw an unwarranted conclusion from the display. The batches may differ greatly in the sample size of the group and the accuracy of the median. In an attempt to improve on the basic plot, we incorporate two additional features into the display, variable width and notches.

The variable width notched box is shown in Fig. 1b. Here the width of the box is proportional to the square root of the sample size. This feature is a valuable aid in preventing the viewer from drawing an inappropriate conclusion from a set of box plots. Significant differences in the sample size are greatly emphasized. Figure 1b also shows a second additional feature, a notch surrounding the median. This notch is proportional to the standard deviation of the median. A constant of proportionality has been chosen in accordance with the recommendation by Tukey,⁷ in which case two medians are significantly different at the 95% confidence level if the notches do not overlap. Tukey suggests this would hold true even in many instances of non-Gaussian data.

In Fig. 2 we see a box plot display of ozone concentration data from six sites in the Imperial Valley. Each batch consists of the maximum value of ozone concentration for each day from December 1, 1976 through December 31, 1977 (396 days). The maximum ozone concentration for each day is the largest of the 96 readings taken throughout the 24-hr day. The 96 readings are filtered by a resistant smoother before the maximum is found.

From the box plot we can instantly get a general picture of the ozone concentration at the six sites. It is obvious that the East Mesa station data is

Fig. 2. Box plot of ozone concentration data from six sites in the Imperial Valley.

different from all other stations both in location and spread.

Box plots can also illustrate the temporal behavior at a single site. In Fig. 3 we see 12 batches, each corresponding to 1 month of daily maximums at the Niland site. The seasonal behavior of ozone concentration is clearly evident. Ozone concentration is maximal during the summer months. We can also see that large median values and large spreads tend to occur simultaneously. Contrast the variation of spread between the winter months and the summer months. This coupling between location and scale usually indicates that a transformation of the data is desirable (statisticians call such transformations, variance stabilizing transformations).

The seasonal behavior for the East Mesa station is illustrated in Fig. 4. We see that ozone concentrations at East Mesa not only tend to be smaller than

Fig. 3. Box plot of ozone concentration data from Niland station for 12 months shows pronounced seasonal effect.

Fig. 4. Box plot of ozone concentration from East Mesa station for 12 months shows lack of seasonal effect. Compare with Fig. 3.

other stations, but fail to exhibit any pronounced seasonal variation. Is it possible that the meteorology at the East Mesa site is also different? Figures 5 and 6 indicate East Mesa is neither hotter nor more windy than other locations. We need to examine the meteorology more closely or look for other factors to explain the reduced ozone at East Mesa.

Fig. 5. Box plot of temperature data from six sites in the Imperial Valley.

Fig. 6. Box plot of wind-speed data from six stations in the Imperial Valley.

Symmetry Plots

Our notions of location and spread are defined for symmetrical distributions. A very useful graphical technique that can be used to assess the symmetry of a set of observations is to plot

$$x_{n-i+1} \text{ against } x_i, i = 1, 2, \dots, n/2.$$

where

$$x_1 \leq x_2 \leq x_3 \leq \dots \leq x_{n-1} \leq x_n$$

are the sorted observations.

If the numbers come from a symmetrical distribution, this plot will look like a straight line with a slope of -1 . In practice, we plot $(x_{n-i+1} - x_i)$ against $(x_{n-i+1} + x_i)$, in which case numbers from a symmetrical distribution will be a straight line with unity slope intercepting the origin.

The daily maximums for Niland station ozone are plotted in Fig. 7. The sharp departure from the unit slope reference line indicates asymmetry. If however, the square roots of the ozone measurements are plotted as shown in Fig. 8, the points now cluster about the reference line. In Fig. 9 the log is used rather than the square root. Comparing Figs. 8 and 9, it is apparent that the square root of ozone is significantly more symmetrical than the log.

The log-normal distribution is generally used to model ozone data; these results show the square root is perhaps a better choice. (Our results are consistent with Bell Laboratories' study of ozone in the Northeast U.S.¹ In Figs. 10 and 11, we have histograms of the raw and transformed Niland ozone daily maxima, the difference in symmetry is striking. The symmetry plot is a much more sensitive indicator than the histogram; departures from a straight line are easy to judge.

Fig. 7. Symmetry display of ozone concentration data from Niland station. Pronounced departure from straight line indicates asymmetry in distribution.

Fig. 8. Symmetry display of ozone concentration data from Niland station. Data have been transformed by taking the square root. Compare with Fig. 7.

Fig. 10. Conventional histogram of ozone concentration data from Niland station. Asymmetry is evident.

Fig. 9. Symmetry display of ozone concentration data from Niland station. Data have been transformed by taking logs. However, data are still significantly asymmetrical. Compare with Fig. 8.

Fig. 11. Conventional histogram of ozone concentration data from Niland station. Data were transformed by taking the square root.

Time-Series Plots

Another useful display in exploratory data analysis is the time-series plot. The structure of the plot can be very critical to the visual impact the plot has. We have to decide whether to connect

the plotted points with lines, or just plot isolated points. We also have to decide how to represent missing data.

For instance in Fig. 12 we plot a resistant correlation coefficient against days for ozone. In this plot, missing days come out as blank gaps. Although

this plot is informative, we can do better. In Fig. 13 we have the same data, except the plotted points are not connected. Instead a vertical line is dropped from the point to the horizontal axis. Days with missing data are plotted as little "whiskers" at the bottom. This plot has a much better visual impact

Fig. 12. Correlation between Niland and Imperial stations for ozone concentration vs time. Individual points are connected by straight lines.

Fig. 13. Correlation between Niland and Imperial stations for ozone concentration vs time. Here, individual points are not connected. Days with missing data are indicated by short "whiskers" at bottom of plot. Compare with Fig. 12.

Fig. 14. Correlation between Niland and East Mesa stations for ozone concentration vs time. Considerable negative correlation is shown. Compare with Fig. 13.

and conveys more information about the data. We see no seasonal variation in the correlation between the Niland and Imperial stations. We note a few "funny" days occur that have negative correlations (lines below horizontal); these days should be investigated separately. In Fig. 14 we see the ozone correlation between Niland and East Mesa. The difference with Fig. 13 is striking. We see more days with missing data and many more days with negative correlation. These time-series plots of interstation correlations indicate that the differences between these stations tend not to show a seasonal variation.

In Figs. 15 and 16 we have time-series plots for ozone daily maxima for the Niland and East Mesa stations. The seasonal variation is quite evident at Niland; the summer months have about twice the concentration as the winter months. We can see that East Mesa ozone concentration exhibits no significant seasonal variation.

Fig. 15. Time-series display of ozone concentration data from Niland station. One vertical line represents 1 day. Days with missing data are indicated by down-pointing "whiskers" at bottom of plot. Note pronounced seasonal effect.

Fig. 16. Time-series display of ozone concentration data from East Mesa station. One vertical line represents 1 day. Compare with Fig. 15. Note especially days with extreme readings. No seasonal effect is apparent.

Resistant Smoothing

The presence of long tailed or "spiky" noise often obscures any obvious pattern in a time series. Conventional techniques smooth sequences by replacing each x_i in the sequence with a weighted average of the points near x_i . These smoothers are linear digital filters. Low-pass filters perform poorly on signals that contain impulsive noise. This is because impulse noise is wideband and the frequencies of the noise extended well into the region of the spectrum occupied by the signal. This is an inherent limitation of all linear filters. We can see why linear filters perform poorly if we think of the noise spikes as outliers and the filter as an estimator based on the mean value. Signal processors sometimes use nonlinear devices, such as clippers, to process signals with impulsive noise.

There are a class of nonlinear smoothers proposed by Tukey that are resistant to "spiky" noise. These smoothers are based on a moving median technique. We will not treat any of the details of these smoothers, merely demonstrate their utility in the exploratory analysis of oxidant data.

In Fig. 17 the solid curve is a diurnal curve of one month σ_θ data. The quantity σ_θ is a measure of variation in wind direction and provides an index of atmospheric turbulence. The solid curve is calculated by first forming a trimmed mean across all the days of the month for each 15-min time interval. Thus the noon value would consist of the trimmed mean of 31 numbers. The 96 trimmed means are then filtered by a nonlinear smoother and the solid curve results. Superimposed on the solid curve is a dotted curve that shows the raw σ_θ readings for the specific day of July 15, 1977. Next we apply the nonlinear smoother to the dotted curve. The result is shown in Fig. 18 where the dotted curve is the smoothed version of the dotted curve in Fig. 17. The structure of this day's diurnal variation is now more clearly revealed. In one plot we can compare a specific day to the average diurnal behavior for the month.

Quantile Plots

There are a number of formal statistical procedures to determine if two sets of empirical observations have been drawn from the same parent population (e.g., Smirnov two-sample test). Unfortunately these procedures usually do not produce a corresponding visual display. A very valuable graphic tool for comparing two univariate sets of observations is known as the quantile-quantile, or "QQ" plot.

Fig. 17. Diurnal curve for sigma theta. Higher values indicate increased atmospheric turbulence. Solid line is smoothed diurnal curve for 1 month; dotted curve is actual readings for indicated days.

Fig. 18. Diurnal curve for sigma theta. Higher values indicate increased atmospheric turbulence. Solid line is smoothed diurnal curve for 1 month; dotted curve is readings after smoothing for indicated days.

Figure 19 shows two cumulative distribution functions (CDF) to illustrate the definition of a QQ plot. Corresponding to any ordinate P there are two quantile values q_x and q_y . If X and Y are identically distributed variables, then a plot of q_y vs q_x would be a straight line with unit slope. If Y is a linear function of X , the QQ plot would still be linear albeit with a change of slope and intercept. A linear relation between Y and X implies that statistically X and Y differ by location and spread. Or equivalently, if there exists a linear transformation one could make on X that would make it identically distributed with Y . The QQ plot should not be confused with a scatter plot. A scatter plot reveals correlation (i.e., linear association)

between X and Y. It is possible for X and Y to have identical distributions, yet be uncorrelated. The QQ plot enables us to compare the entire set of numbers Y against the entire set of numbers X.

We now apply QQ plots to compare ozone concentrations at different stations. In Fig. 20 we have a QQ plot of ozone daily maximums at Niland vs Imperial. A reference line of unit algebraic slope is indicated. It is immediately evident that the ozone concentrations at Niland are consistently higher than Imperial. Note especially departures at the high

Fig. 19. A plot of $q_y(p)$ vs $q_x(p)$ for $0 < p < 1$ produces a quantile-quantile plot.

Fig. 20. Quantile-quantile plot comparing ozone concentrations at two stations in the Imperial Valley. Note significant departure from indicated unit slope reference line at higher values.

values. In Fig. 21 the ozone daily maximums for Niland are compared to East Mesa. We see a very significant departure from the reference line indicating the two sets of measurements have very different distributions.

Fig. 21. Quantile-quantile plot comparing ozone concentrations in the Imperial Valley. Note significant departures from reference line at majority of levels.

Conclusions

High level ozone concentrations have been found in a number of non-urban regions in the United States. The Imperial Valley is another location with this characteristic. For about 25% of the days under study, the maximum ozone concentration exceeded the federal standard of 80 ppb. Levels as high as 220 ppb were found for some days. These findings are at variance with the concept that ozone is a direct result of emissions by man-made sources, particularly the internal combustion engine.

Our exploratory analysis has revealed part of the spatial temporal behavior of ozone concentrations. We have found that one site, East Mesa, stands out as being very different for all modes of comparison. The ozone levels at East Mesa are closer to what we would expect of a rural region.

At this point we have two major hypothesis. The high ozone levels are the result of direct transport from populated areas, such as Mexicali to the south, or San Diego and Los Angeles to the north and west. There is evidence of ozone transport over long distances in the northeast.

The second hypothesis is that the high ozone levels are the result of agricultural activities in the Imperial Valley. It is very significant that the East Mesa location is the only site not surrounded by major agricultural development.

Further study into the meteorological structure of the valley would be necessary to investigate the first hypothesis. Correlations of wind direction with ozone levels would be valuable. Correlation of ozone levels with some measure of agricultural activity (such as water usage) could shed light on the second hypothesis.

The exploratory methods described here are not limited in application to air quality data. They are particularly valuable when physical models are not available as a guide and when the data under investigation may not be well behaved.

References

1. D. F. Andrews, et al., *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, NJ (1972).

2. R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots" *The American Statistician* **32** (1) (February 1978).
3. W. S. Cleveland, B. Kleiner, J. E. McRae, and J. L. Warner, "Photochemical Air Pollution: Transport from New York City Area into Connecticut and Massachusetts," *Science* **191**, 179 (1976).

Further Readings

- W. S. Cleveland, B. Kleiner and J. Warner, "Robust Statistical Methods and Photochemical Air Pollution Data" *APAC Journal* **2b**(1) (1976).
- Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations* (John Wiley, New York, 1977).
- National Academy of Sciences, "Ozone and Other Photochemical Oxidants," *Committee on Medical and Biological Effects of Environmental Pollutants*, Washington, D.C. (1977), pp. 13-44.
- Tukey, J. W., *Exploratory Data Analysis* (Addison Wesley, New York, 1977).