LBL-8874 c.2
UC-92

AN AGGREGATED VECTORIAL MODEL OF PETROLEUM FLOW IN THE UNITED STATES

V. V. Krishnan and D. F. Cahn

March 1979

Prepared for the U. S. Department of Energy
under Contract W-7405-ENG-48

LBL-8874 c.2

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal 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. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-8874

AN AGGREGATED VECTORIAL MODEL
OF PETROLEUM FLOW IN THE UNITED STATES

V. V. Krishnan and D. F. Cahn

March 29, 1979

Information Methodology Research Project
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

i d a d
INFORMATION &
DATA
ANALYSIS
DEPARTMENT

ABSTRACT

An aggregated material-flow model is proposed for crude oil and its derivative products. The purpose of this model is to isolate stages in petroleum flow where material conservation is expected from those where volumetric or identity changes can occur, and to identify generic properties of petroleum and petroleum products that would assist in effective data validation. The model also provides a structural framework for organization and consolidation of the various databases related to petroleum, and serves as a guide for analysis and enumeration of explicit semantic data interrelationships. The model is amenable to expansion into both transactional and more disaggregated representations.

In the present study, the material-flow model was intended as a preliminary step toward a coherent and comprehensive data structure to support monitoring, forecasting, and regulatory efforts in the energy field. The model is developed in the abstract; no attempt has been made to test it using explicit data.

I. INTRODUCTION

In the wake of the 1973 Arab oil embargo and in the face of budding energy crises, Congress created the Energy Information Administration to organize information and data in the energy field so that better assessments of U.S. energy needs and resources could be made than had heretofore been available [1]. The EIA inherited a wealth of data originally collected by several federal agencies under various legislative mandates and organized into approximately 230 data bases and forecasting models, each addressing specific issues in the energy field. The responsibilities of the EIA are aimed primarily at the global goal of support for energy policy decisions. The EIA is, therefore, required to validate and organize the existing data; further, it is required to identify areas in which the existing data are insufficient to provide a clear picture of the energy balance and in which new data should be collected.

The existing data systems were developed independently, without meaningful standards for information quality or uniformity. Because the volume of existing data is immense and of unknown validity, coherent integration into a single information system is a monumental task that involves validation both within and across the existing databases. It is not at all clear that integration of disparate databases is preferable to creation of a new, unified system *ab initio*. However, regardless of the approach to be taken, there are two essential facets of the design:

(1) A clear definition of the information components of the system down to the lowest level (data elements) is required. This is a microscopic task, involving analysis (in the present systems) of about 9600 variables (with as yet undetermined independence) and their interrelationships. Such an analysis is the basis for the semantic relations either in a unified database or in the coordinated usage of the existing databases.

(2) It is necessary to develop a framework or data model with a structure that provides for logical organization of the data, that permits the extraction of all information derivable from the data, and that makes it possible to identify what additional data are required to answer a specified query.

The data model defines semantic relations between logical system components in the unified database.

The present study addresses the second question, the development of a data model capable of representing accurately the relationships between observable quantities. In the domain of energy supply, the problem is equivalent to the analysis of physical product flows (and of product transactions) through the system. Some efforts have been made in this direction, most notably in the PIES model [2], although the high levels of aggregation addressed by such models preclude their application to the data structural analysis necessary to extract much of the information contained in the system and routinely required to answer questions about validation, short-term policy options, and regulation. In order to insure the completeness and consistency of any developed data structure, it is necessary to approach the problem from high levels of aggregation and work downward toward greater resolution; at the same time, however, it is necessary to establish a clear data path between the highly aggregated variables and the base-level raw data.

The purpose of this study is to formulate such an aggregated, but explicitly linked, structure for a sample segment of the energy system -- that relating to crude oil and its derivative products. The methodology employed is applicable generally to the organization of energy data; we have used the oil system merely as a sample environment. A vectorial flow structure is developed that isolates transmission vectors on a product basis from processing and interchange that takes place in the system. By isolation of those stages at which identity and volume changes can occur (such as refining) from those at which material identity and volume are expected to be conserved (such as transportation), it is possible to select generic data nodes at which product monitoring can give validation crosschecks.

Initially, the product flow structure was predicated on use of the entire United States as the control volume, but, as insight has been gained, it has been possible at least to postulate the eventual regional or statewise breakdown. When the product flow structure is developed to a sufficiently high resolution, a transactional flow structure may be implemented in superposition, providing a logical data structure that is capable of monitoring many aspects of the petroleum energy supply system.

It must be emphasized that the work reported here was conducted in a limited time frame and represents only a first cut at the problem. The aim was to develop a general and coherent model data structure, not to deal exhaustively with all the specific semantic issues associated with petroleum data itself. Consequently, such specifics as the average purchase price equation mentioned in Section II.D. below should be interpreted (as they were intended) as one candidate examples that are consistent with the model, but have associated advantages and disadvantages. Consideration of optima among such specific semantic choices is explicitly postponed to later efforts.

II. CHARACTERISTICS OF THE MODEL

The aggregated petroleum flow model developed here is based on a review of several existing petroleum-related models and has been particularly influenced by the approach of the PIES model [2]. The underlying methodology of our approach is to view the flow of petroleum through the economy as a sequence of logically coherent stages and processes through which petroleum passes between its extraction from the ground (or its importation into the U.S.) and its use by the consumer. During its passage through these stages and processes, the petroleum undergoes transformations both in volume and in chemical composition (from various types of crude oil to various types of petroleum products), transportation between various geographical locations, and fluctuations in price and the characterization of its intended use.

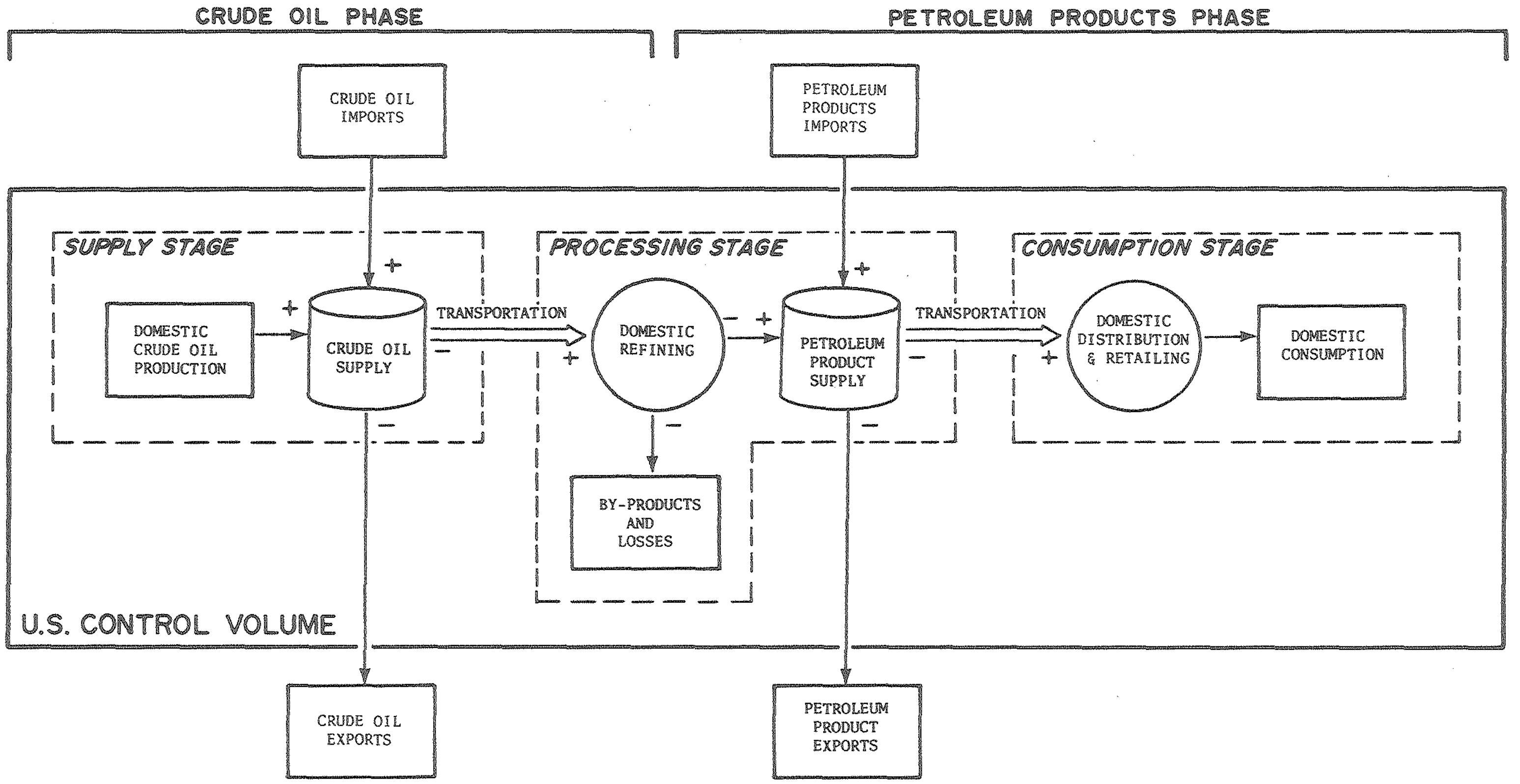
In our model, oil in the system is considered to exist as a set of quantized units or "packets". Each oil packet has associated with it a set of attributes that, when organized into a vector, provides a description of the current "state" of the packet sufficient to characterize it in all aspects relevant to the intended use of information about it. In particular, the attribute set must be sufficient to answer oil-related questions concerning, for example, types (by chemical composition, tier, etc.) of petroleum in the system, present locations and volumes of oil or petroleum products, volume of oil (or product) in storage and volume in transit, geographic source and destination, transportation medium, and current price at any given stage. The packets are considered ultimately to represent individual oil shipments or consignments; the packets and attributes together determine the resolution grain of the model, and thus the combinatorial questions it can answer. The packet view applies in general over many aggregation levels and control volumes; thus, in the attempt to organize our view of these systems, it makes sense to use these representations, even if at present we do not have access to an organized cache of high resolution data.

The various stages, processes and attributes described below are meant primarily to convey the basic principles of our approach and should not be construed as a finalized description of the actual system. We have, however, chosen them with sufficient care to allow us to expand or contract the scope and detail involved in any single stage or attribute without unduly interfering with the remainder of the model.

A. Stages of the Model

The petroleum flow model presented here (Figure 1) is organized at its most general level into three stages - Supply, Processing and Consumption - following a scheme similar to that of the PIES model, which also subdivides petroleum flow into three stages (Supply, Processing and Demand). We have emphasized "consumption" rather than "demand" to indicate that the model represents actual use of petroleum rather than predictions. In our model the three major stages are connected by unidirectional transportation links (denoted by thick arrows in this and subsequent figures). Additional physical transportation of crude oil and petroleum products may occur within the individual stages, but at the highly aggregated level of Figure 1, this is minor.

As Figure 1 shows, the Supply stage reflects total crude oil available for domestic use in the mainland U.S.A. as the sum of domestically produced crude and the crude oil imported from abroad minus the amount of crude oil exported from the United States.



XBL 793-8947

FIGURE 1. Major flows in the U.S. petroleum system, simplified form. Three major functional stages and two material phases as shown. Thick arrows denote physical transportation. Heavy boundary indicates control volume boundary for the U.S.

The output of the Supply stage drives the Processing stage, in which crude oil is converted into gasoline and other petroleum products. Petroleum product imports form an additional source that feeds the system here, and petroleum product exports an additional drain.

The Processing stage in turn drives the Consumption stage. The Consumption stage can in fact be further subdivided into numerous sub-stages (as is done in most econometric models), but we have refrained from doing so in the initial aggregated flow model.

In addition to functional partitioning into stages, the model distinguishes two material flow phases, one for crude oil and the other for oil products, as denoted by the horizontal brackets at the top of Figure 1. The crude oil and oil products phases are considered to operate quasi-independently. The connecting link, the refining process, requires an additional model. However, given the process latitude available in refining and the many economic as well as chemical and physical factors that affect the product mix produced at each refinery, refining models adequate for material accounting purposes are not presently available. Chemical reaction stoichiometry, while a limiting factor, is modifiable by processes such as catalytic cracking and blending, and individual refineries vary the products they produce from each variety of crude over a broad range dependent on anticipated supply and demand in the regions they service. In the case of national or international producers, product decisions at individual refineries under their control may be

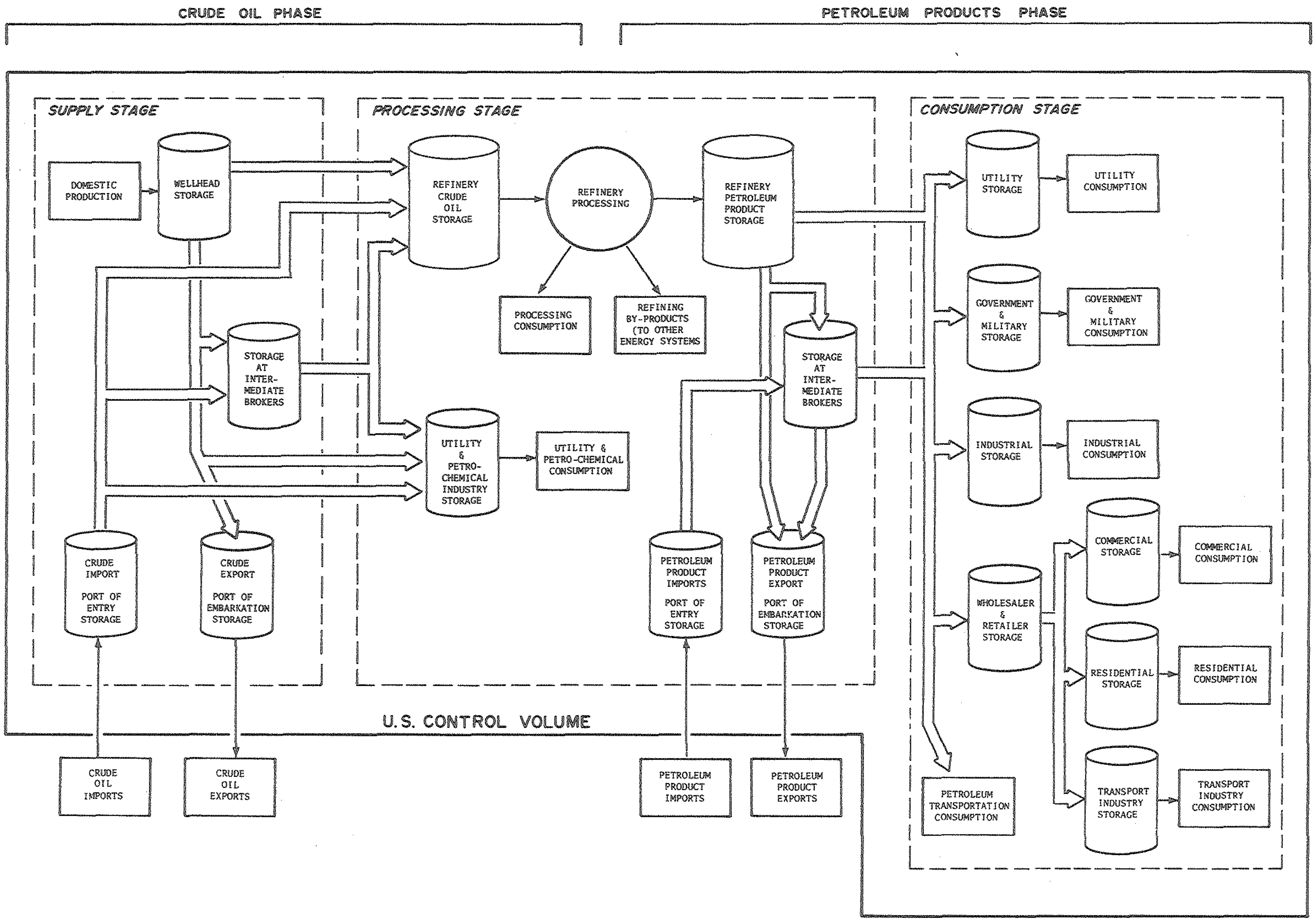

CRUDE OIL PHASE
PETROLEUM PRODUCTS PHASE
SUPPLY STAGE
PROCESSING STAGE
CONSUMPTION STAGE
DOMESTIC PRODUCTION
WELLHEAD STORAGE
REFINERY CRUDE OIL STORAGE
REFINERY PROCESSING
REFINERY PETROLEUM PRODUCT STORAGE
PROCESSING CONSUMPTION
REFINING BY-PRODUCTS (TO OTHER ENERGY SYSTEMS
STORAGE AT INTER-MEDIATE BROKERS
UTILITY & PETRO-CHEMICAL INDUSTRY STORAGE
UTILITY & PETRO-CHEMICAL CONSUMPTION
STORAGE AT INTER-MEDIATE BROKERS
CRUDE IMPORT PORT OF ENTRY STORAGE
CRUDE EXPORT PORT OF EMBARKATION STORAGE
PETROLEUM PRODUCT IMPORTS PORT OF ENTRY STORAGE
PETROLEUM PRODUCT EXPORT PORT OF EMBARKATION STORAGE
UTILITY STORAGE
UTILITY CONSUMPTION
GOVERNMENT & MILITARY STORAGE
GOVERNMENT & MILITARY CONSUMPTION
INDUSTRIAL STORAGE
INDUSTRIAL CONSUMPTION
WHOLESALER & RETAILER STORAGE
COMMERCIAL STORAGE
COMMERCIAL CONSUMPTION
RESIDENTIAL STORAGE
RESIDENTIAL CONSUMPTION
TRANSPORT INDUSTRY STORAGE
TRANSPORT INDUSTRY CONSUMPTION
PETROLEUM TRANSPORTATION CONSUMPTION
U.S. CONTROL VOLUME
CRUDE OIL IMPORTS
CRUDE OIL EXPORTS
PETROLEUM PRODUCT IMPORTS
PETROLEUM PRODUCT EXPORTS

FIGURE 2. Major flows in the U.S. petroleum system. Cylinders denote matrices of storage facilities, rectangles denote material sources and sinks, circles denote material transformation sites. Other conventions as in Figure 1. (All transportation stages have associated fuel usage.) XBL 793-8945

optimized on bases ranging from international to local; as these corporate strategies are, for all practical purposes, inaccessible to regulatory agencies, attempts to model the refining process itself are presently untenable. By accounting for crude oil and petroleum products as separate phases, one can sidestep the latitude and relative unpredictability of the refining process and thus preserve the accuracy and usefulness of the model. The pre-refining material, crude oil, is traced until its consumption at the refinery inlet, and the post-refining materials, petroleum products, are traced from their source at the refinery outlet. The process connection is recognized and acknowledged, but is specifically eliminated from the model pending a more formalizable characterization.

The functional stage and material phase subdivisions of the Figure 1 model are carried over to a more detailed presentation in Figure 2, as are the use of thick arrows to denote physical transportation, cylinders to denote storage, and rectangles to denote material sources and sinks. Consequent to the arguments above, the first phase of the Figure 2 model traces the flow of crude oil from the point of drilling or the point of importation to the point at which it accumulates in refinery storage tanks awaiting processing or in an embarkation port awaiting exportation. The second phase deals with processed products from the time they flow into refinery or port of entry storage tanks to the point at which they are sold and transferred to the storage tanks or other facilities provided by the consumer.

Both model phases consider flows within a control volume limited to the United States. The control volume bounds are indicated in Figures 1 and 2 by heavy lines. Imported oil is traced only from the point at which it crosses the U.S. control volume and goes into storage at a U.S. port of entry, and exported oil only to storage at U.S. port of export. One might reasonably extend this model to the country of origin in the case of imports and to the country of destination in the case of exports, but this was not done in the present model because, on an ongoing basis, the requisite transportation data may not be obtainable within the authority bounds of U.S. agencies unless the oil is transported by U.S.-based carriers, and thus may be expected to be generally incomplete.

The material flow and handling stages shown in Figure 2 form a prototype interaction net sufficient to mirror the instantaneous state of the entire system. As stated earlier, the first round effort reported here was aimed at finding principles adequate and efficient for representing and interlinking data in systems such as this one, and was conducted without extensive analysis of the existing data. Consequently, we do not claim absolute completeness at this stage for the interaction model of Figure 2, but present it instead as a first order approximation model for the system, subject to further refinement but adequate for us to test the principles we believe appropriate for designing and implementing an information system dealing with energy data.

Each storage element (cylinders) of the Figure 2 model (and each transportation element of the model (thick arrows)) is in fact composed of a data matrix containing specific instances of the generalized unit named (or connecting named elements, in the case of the transportation links). The cylinder representing "Wellhead Storage", for instance, is a matrix containing data on all the storage facilities at well sites throughout the country; for each such site, information on volume, location, owner, etc., is organized, for each distinct type of crude oil kept separately in storage, into a vectorial representation. The matrix ensemble of these "attribute vectors" gives a picture of the entire wellhead storage of crude presently within the country, regardless of location; i.e., wellhead storage is organized into a generic class.

Crude oil can flow from wellhead storage to refineries, to intermediate brokerage and resale, to export, or to the utilities and petrochemical industry; by isolating the generic classes of storage that crude may undergo (as described above), the number of general interaction path types connecting elements is shown to be small. Admittedly, the actual number of site-to-site paths is much larger than the four general ones noted here, but organizing them on a generalized basis allows the interaction links themselves (in this case involving physical transport) to be represented in a homogeneous matrix covering the entire country. (Source and destination simply become two of the attribute data values associated uniformly with each interaction link.) The uniformity of the representation is, of course, a key requirement in applying computer information processing procedures advantageously.

The remainder of Figure 2 can be analyzed in the light of the discussion above. In all cases, the storage elements and their interlinks are generalized, and are considered to contain all specific instances of their data types throughout the control volume. Thus the wellhead storage block can contain entries that feed both domestic refining and exportation, while the refineries draw on both domestic wellhead storage and importation; as in the real situation, there is no contradiction here, since, for example, Alaskan oil may not be transportable to New England, and is thus exported to Japan, while New England must import from the Middle East to meet its needs.

At any block, material inflows and outflows take place unidirectionally, as shown by the arrows. A summation over all flows so represented provides a continuity equation for the block, and, barring sourcing or consumption within the block (intentionally isolated throughout the model and restricted to the rectangular blocks shown in Figure 2), should give a conclusive picture of all state changes affecting the block. This continuity principle, carried forth over the entire system, provides a powerful mechanism for data validation.

In the data model, each class of unit (for example, wellhead storage) is represented by a matrix whose elements are subunits of the class. The size of the subunits determines the resolution or grain size. For example, an element of the matrix representing wellhead storage could represent all such storage belonging to a company, the portion located within a specified state, a specified tank farm, or even a single tank. The usefulness of the model (and perhaps the cost of collecting and maintaining the data) will increase with increasing resolution; however, the properties of the model, as described below, are independent of grain size. In particular, the form of a data transformation representing oil flow (or representing a purely financial transaction) is independent of the level of aggregation.

The matrix representation of an oil storage unit and the transformations representing oil flow are discussed in Section II.D. Before considering them, it is necessary to define the variables that describe the fundamental unit of oil flow and the characteristics of oil-handling facilities.

B. Packet Attribute Vectors

Each flow packet in the system is represented by a vector of attribute values describing its contents. The twelve variables defined below are sufficient to characterize a packet in any stage or phase of the system. Generally, the attributes reference dimensions of the element blocks (to be presented in Section D below), as well as addressing values within the transition vectors.

While potentially a large volume of data would result from the twelve data elements multiplied by acceptance (receiving) and release (shipping) of all packets in the system, the actual reporting burden on respondents should be considerably less than at present because the data are, perhaps with minor modification, raw entries from the shipping/receiving forms normally used in the course of business. The use of raw data in this manner should actually ease the regulation burden on the reporting companies (who would otherwise have to collect it individually anyway in order to generate the macroscopic totals presently required), while dramatically enhancing validation paths within the data and limiting the effects and propagation of reporting bias.

The variables in the attribute vectors have several general forms. Some, such as volume (Q), are continuously variable numbers; others, such as price tier of crude (T), are discrete variables which can take on only a finite set of values; still others, such as the physical geographic location (I) are linguistic and are treated as only alphabetic or logical units.

Explicit values of attribute variables are appropriate in different measures to different stages of the model. For instance, the price tier attribute (T) is applicable only to crude oil and has no significance for petroleum products. This attribute will have non-zero entries in the supply-processing phase but will have no numerical values in the processing-consumption phase of the model, where the material flow is made up only of petroleum products and not crude oil.

The attributes presently considered relevant to the petroleum state vector are briefly described below.

$$STATE_{packet} = f(Q,A,T,P,I,J,K,W,D_I,D_F,x)$$

Q: Volume of oil being transported or stored. (As used below, oil represents both crude oil and petroleum products in the vector.)

A: Type of oil based on chemical composition. (It is not clear, at present, what specific details on chemical composition are needed for a clear classification of this attribute. It is possible that the attribute may be divided into other sub-attributes in a later version of the model.) In an extremely aggregated version we can classify crude oil as sweet, sour, etc. and the products as light, middle and heavy distillates. On the other hand, one might follow the classification scheme used in the World Energy Model [3] and classify the unrefined petroleum (crude) into 52 different types based primarily on the oil-field from which the crude was extracted. (EIA also has a similar classification for imported crude oil based on the country of origin.)

It may also be desirable to classify the petroleum products by chemical composition in a highly disaggregated form as is done presently by some of the EIA data collection systems. Thus the entries for the attribute A for petroleum products could be any one of the following: Motor gasoline, Aviation gasoline, Naphtha-type jet fuel, Kerosene-type jet fuel, Kerosene, Distillate fuel oil, Fuel oil No.4, Unfinished oils, etc.

T: Oil price-tier per the U.S. Federal price-tiering structure. At present the price-tier structure applies only to crude oil, and therefore the only entries for this component of the attribute vector will be in the supply-processing stage. Currently, "Old oil" and "New oil" have different price ceilings although their chemical composition may or may not be different. One would assume that the Congress may modify this structure further and that this classification is of considerable significance to the regulatory functions of DOE. At least two existing data-collection systems, the Crude Oil First Purchaser system and the Crude Oil Entitlements system, presently gather data on this attribute.

P: Price per unit volume actually paid for the packet by the company reporting the transaction. Even within a single stage, oil price changes substantially depending upon transportation undergone by the oil, storage costs, the number of brokers involved in the various transactions through which it has passed, economic conditions, etc. This data is not presently available for all stages of the model although some data systems collect it for specific products. It is not clear how this information can be obtained for all transactions without a specific legal mandate, since companies generally guard it jealously, but it is clear that such information is vital for regulatory functions.

I,J: Both I and J denote geographical regions associated with either transportation or storage of a given packet of oil. When the reported datum refers to an oil packet in transit, I represents the origin of the shipment and J represents the destination. When the reported transaction concerns oil in storage, I represents the location of the storage facility and there is no entry for J.

I and J are coded to represent different geographical regions. The current practice followed by most EIA systems is to code them by PAD regions [4] but this is too highly aggregated for many purposes. Since all forms on which data are reported require the respondent to supply state and Zip Code, coding regions by State or County would pose few additional problems. This would also allow data collected by State and County agencies to be incorporated into the model.

K: Transportation mode used for a particular shipment. There are five major modes of transport currently in use for oil: pipelines, waterways, coastal tankers, railroads and trucks.

W: Owner of the packet. It is entirely possible (1) that neither shipper (I), receiver (J), nor transportation agent (K) may own the oil they are handling, and (2) that an oil packet can change hands without being moved at all. Further, (3) there is considerable latitude in the reporting of ownership transactions (for example, with a shipper reporting a packet sold as of date of order receipt and a receiver reporting it purchased as of date of payment, possibly two months later). Given these situations, it is necessary to consider packet ownership as distinct from its current location or destination.

C: Consignee of the packet. Necessary for the same reasons as W. Further, since a packet may go through several sets of (I,J,K) before delivery to its ultimate consignee C, there is a further impetus to defining a separate variable.

D_I,D_F: Initial and final dates of packet transition across a given port. If, for example, it takes three days to pump a 200,000 barrel gasoline consignment into tank cars for shipment from refiner I to distributor J, then I's outlet port is involved with the shipment from (D_I) to (D_F = D_I + 3). Since, especially in the case of synchronous reporting intervals for cumulative totals at various facilities, this represents a reporting error band, it is useful to know what state each shipment is in while it is in transition. These dates are also useful in verifying identity of shipments being traced.

x: A local shipment identifying index. This need not be assigned systemwide (which would be burdensome), but merely serves to match up a release event at the outlet of one block with an acceptance event at the inlet of the block immediately downstream. Ideally, could be a shipping document or invoice number.

C. Facility Attribute Vectors

Beyond the primary attributes associated directly with individual oil shipments or packets, certain secondary attributes have come to light associated with the physical facilities that handle the oil. In the block representation, the facility variables generally act as background constraints and modifiers on the packet-variables handled in the block model. Most of these are transportation-related, but at least one, Capacity (C), is relevant both to transport and storage. In both flow states, C gives the maximum storage available at the facility. By contrast, Q, the packet volume, represents actual volume of a shipment. Thus, for a storage tank-farm, the sum of Q's-in minus the sum of Q's-out (a conservation equation, as detailed under

"Validation" below) gives the volume actually present at the facility at some instant, but C gives the volume the tank farm could contain if it were full. Similarly for transport, C would denote the total capacity of a pipeline or a fleet of trucks, regardless of whether or not it is full at a given moment. Clearly, regulation strategies such as stockpiling oil in anticipation of an impending supply shortage or embargo depend on maximum capacity of the system, and knowledge of these maxima greatly enhances such emergency decision-making.

Facility-related variables identified to date are as enumerated below, and it is anticipated that they have major effects on decisions relating to the oil packet variables, most notably I, J, and K. They are isolated from the packet variables by their much less frequent change, and this greater stability makes it computationally efficient to maintain them in a separate information space.

$$\text{STATE}_{\text{facility}} = g(D,TC,t,CY,C)$$

D: Distance from I to J using mode K. This would refer to the shortest possible distance when there are several possible routes from I to J using K. (Note: The database for D can be in the form of an I,J matrix where the entries represent the distances. This is currently being used in the PIES model on a limited scale.)

TC: Tariff cost data for transportation mode K from location I to location J. This will be a fairly complex database since costs/tariffs depend, among other things, on such factors as volume involved, time of the year, priority ratings, etc. (Note: This attribute would be vital if our data-structure is to be interfaced with an econometric forecasting model.)

t: Time taken for transporting crude oil or petroleum product from I to J using mode K. This would be an average value of time and can be stored in a matrix form. (Note: This attribute will be particularly particularly fuzzy in nature.)

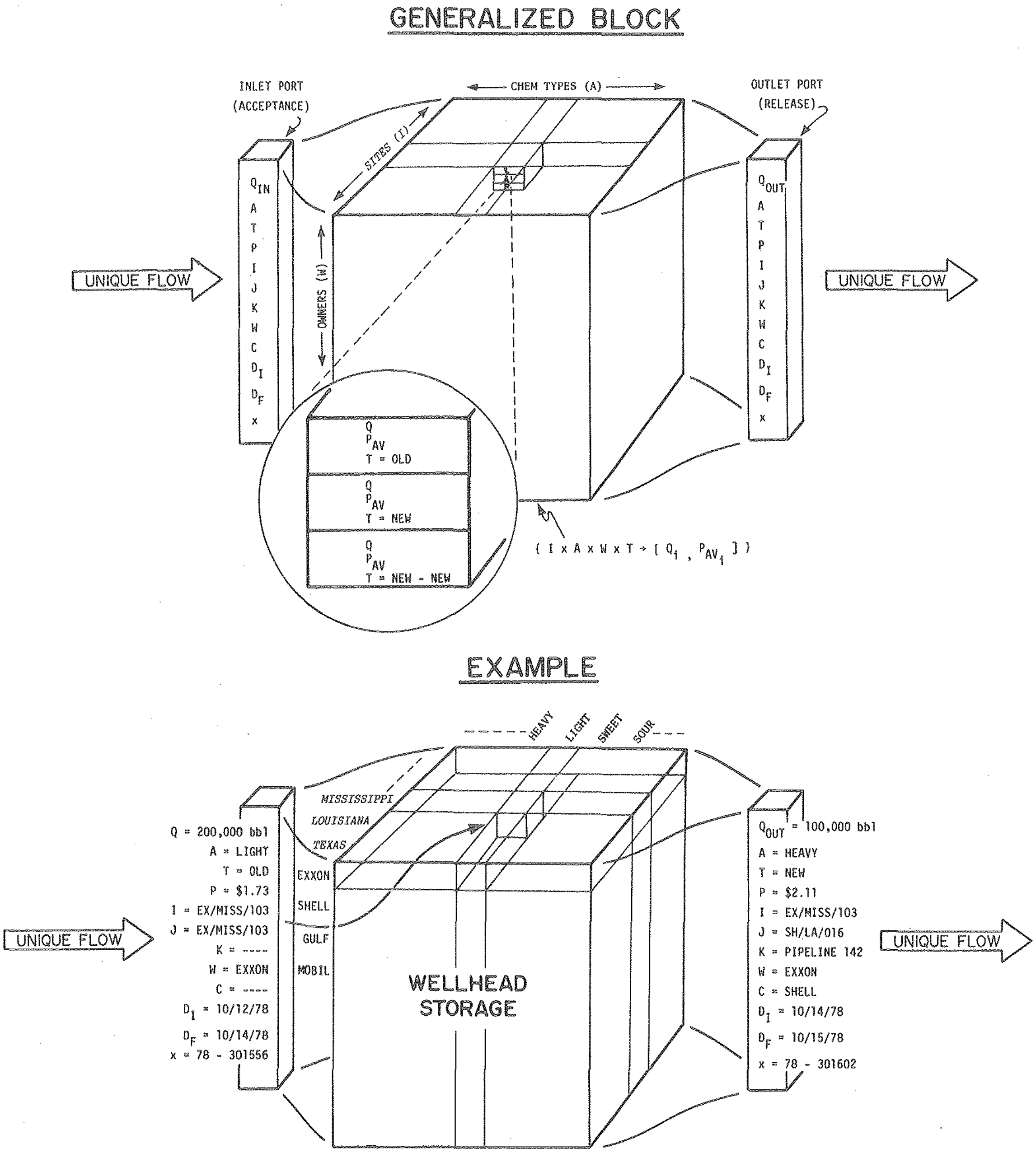
CY: This attribute represents the average cycle time for transportation mode K, for example, the turnaround time for a ship. This information is not absolutely necessary for the model, but will be very useful if our data-structure is interfaced to a policy model.

C: Facility capacity. For storage facilities, the maximum volume or mass storable. For transportation, the total contained in a pipeline, truck fleet, etc., when operating at optimum efficiency. Note that, since products generally cannot be mixed, this variable in fact must be quantized: for example, for a tank farm, C is really a sum of the capacities of individual tanks, any of which can contain only a single homogeneous product at any moment. Thus, C has a bearing on primary attribute A, and there is, instantaneously, a maximum C for every chemical type A.

D. Structural Element Blocks

Consistent with the initial modeling goal of isolating the various material processing steps in the petroleum energy system and enumerating the flow channels connecting them, the process model of Figure 2 can be further subdivided into blocks (Figures 3 and 4), and the blocks then "plugged" together via explicitly constrained interconnections ('ports'), as shown in Figure 5. Material flow between blocks is universally characterized as occurring in packet units, and, as mentioned earlier, each packet is completely specified by its associated attribute vector. The blocks themselves are, in fact, data matrices, and contain cumulative information on material in the system arranged by appropriate attribute dimensions of the packet vectors.

(a) STORAGE BLOCK



GENERALIZED BLOCK

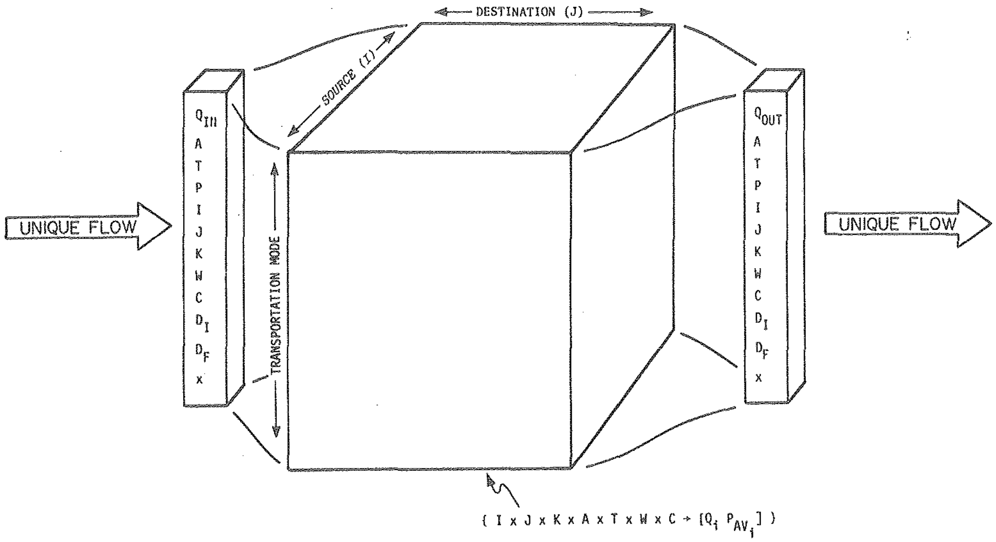
EXAMPLE

FIGURE 3. Matrix block representations for model elements. Generalized forms and examples for each of the five block types that, together, completely represent all portions of the petroleum model:(a)- Storage; (b)- Transportation; (c)- Source; (d)- Sink; (e)- Processing (refining). Vectorial packet input/outputs depicted.

XBL 793-8940

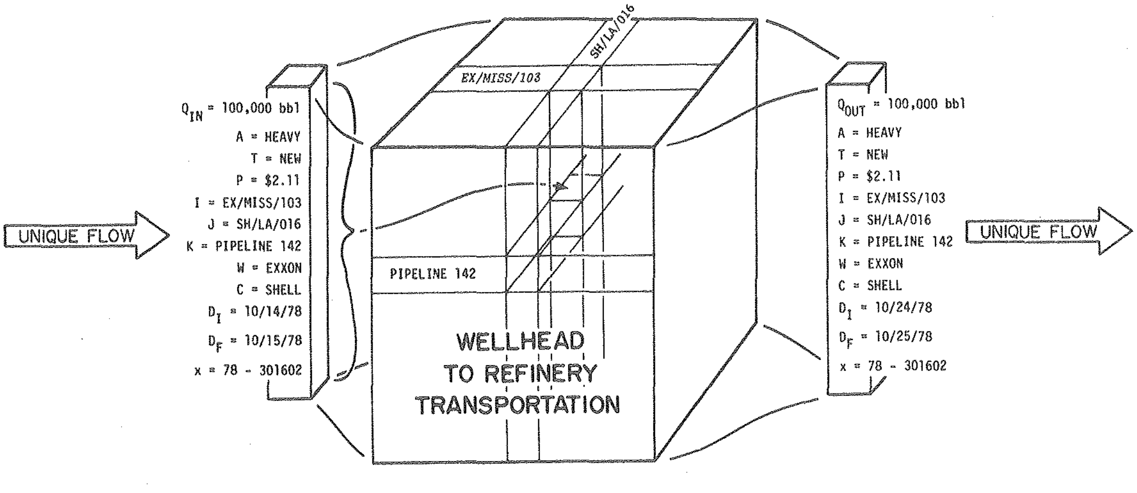
(b) TRANSPORTATION BLOCK

GENERALIZED BLOCK



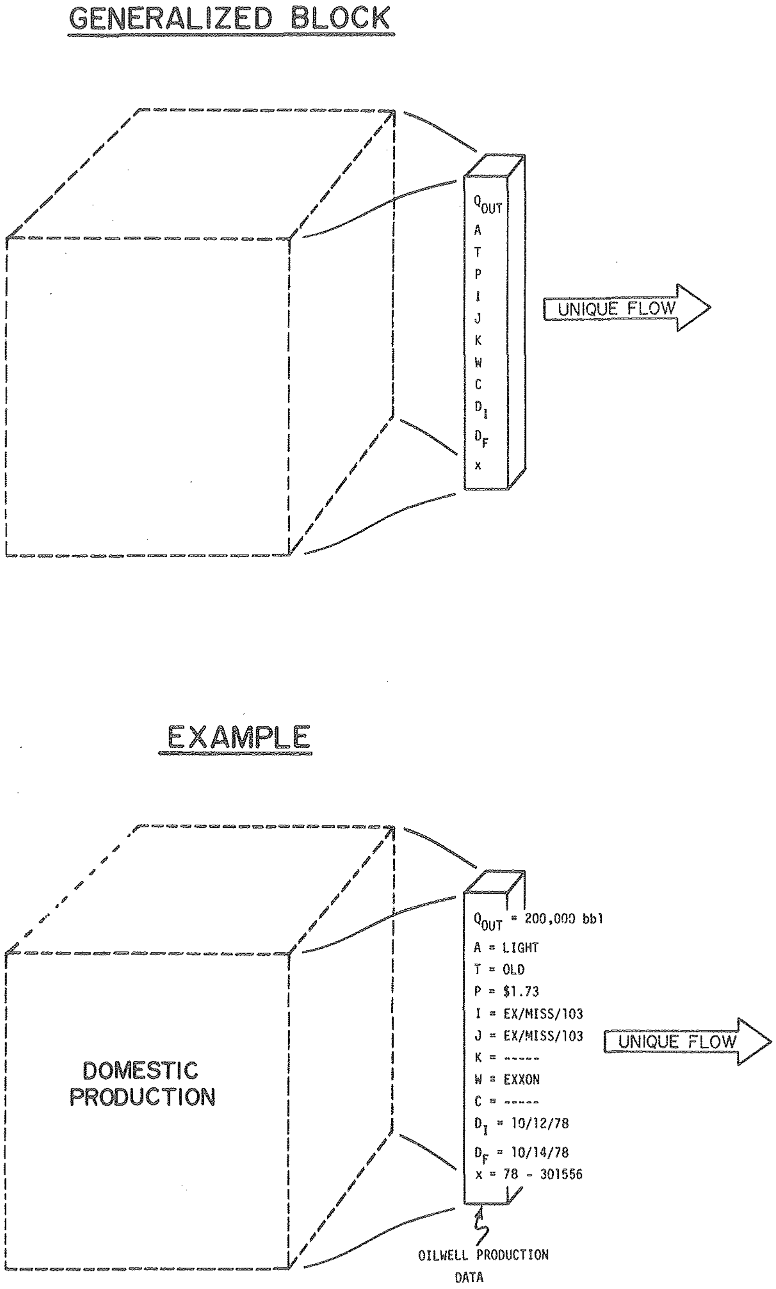
DESTINATION (J)
SOURCE (I)
TRANSPORTATION MODE
UNIQUE FLOW
Q_{IN} A T P I J K W C D_I D_F x
Q_{OUT} A T P I J K W C D_I D_F x
UNIQUE FLOW
{ I x J x K x A x T x W x C → [Q_i P_{AV_i}] }

EXAMPLE



XBL 793-8941

(c) SOURCE BLOCK

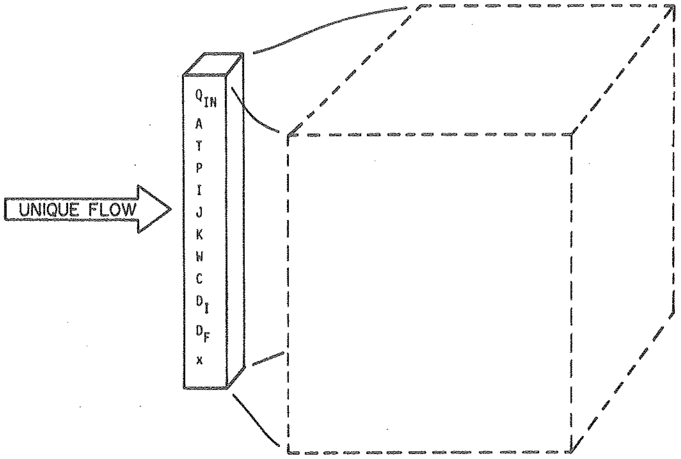

GENERALIZED BLOCK

EXAMPLE

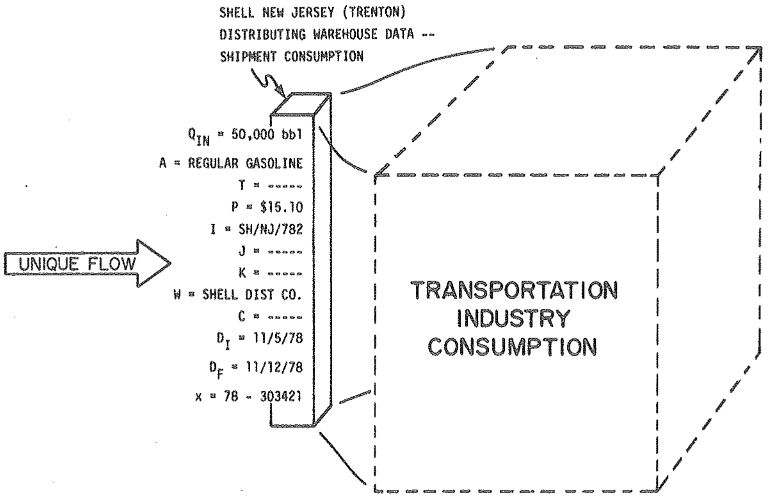
XBL 793-8942

(d) SINK BLOCK

GENERALIZED BLOCK



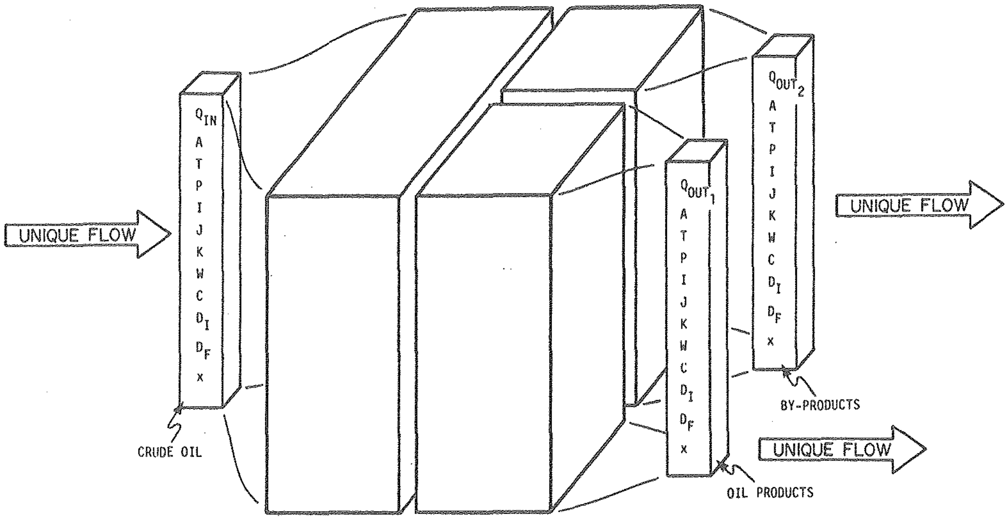
EXAMPLE



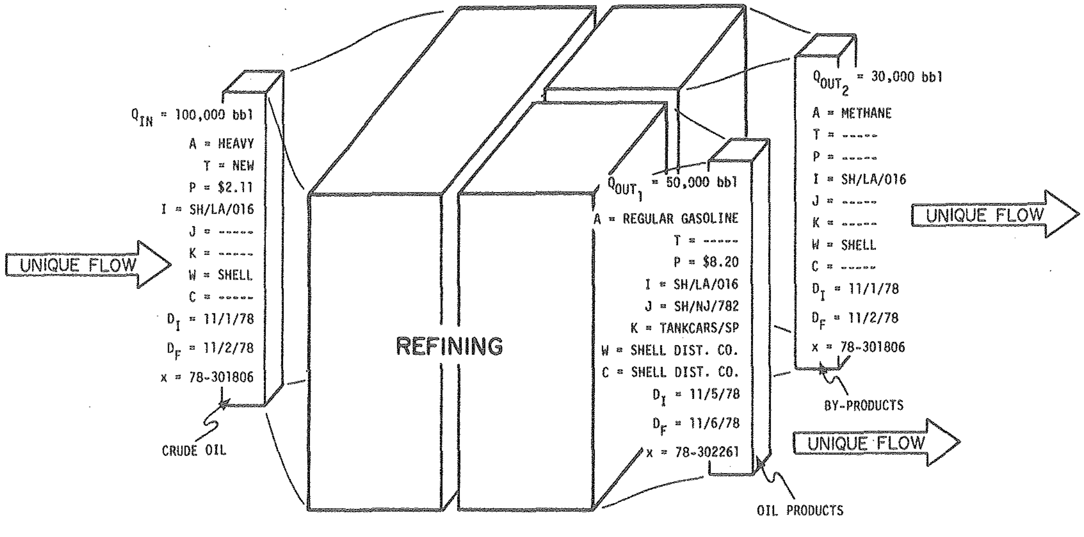
XBL 793-8943

(e) PROCESSING (REFINING) BLOCK

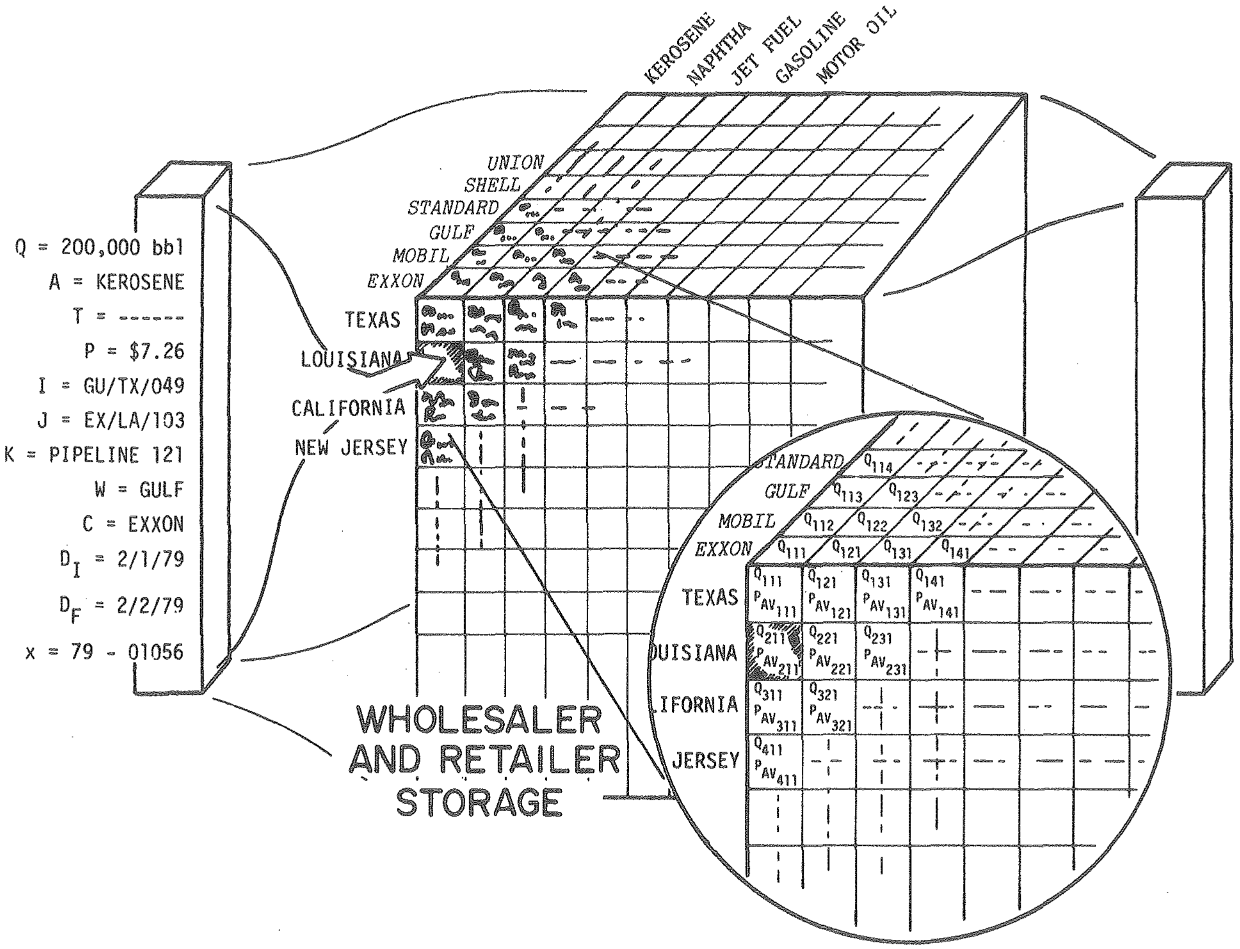
GENERALIZED BLOCK



EXAMPLE



XBL 793-8944



XBL 793-8946

FIGURE 4. Effects of a petroleum packet received through the inlet port of a block. Received packet is routed to the appropriate element of the storage matrix, and modifies its entries for volume Q and average price P_{AV} accordingly.

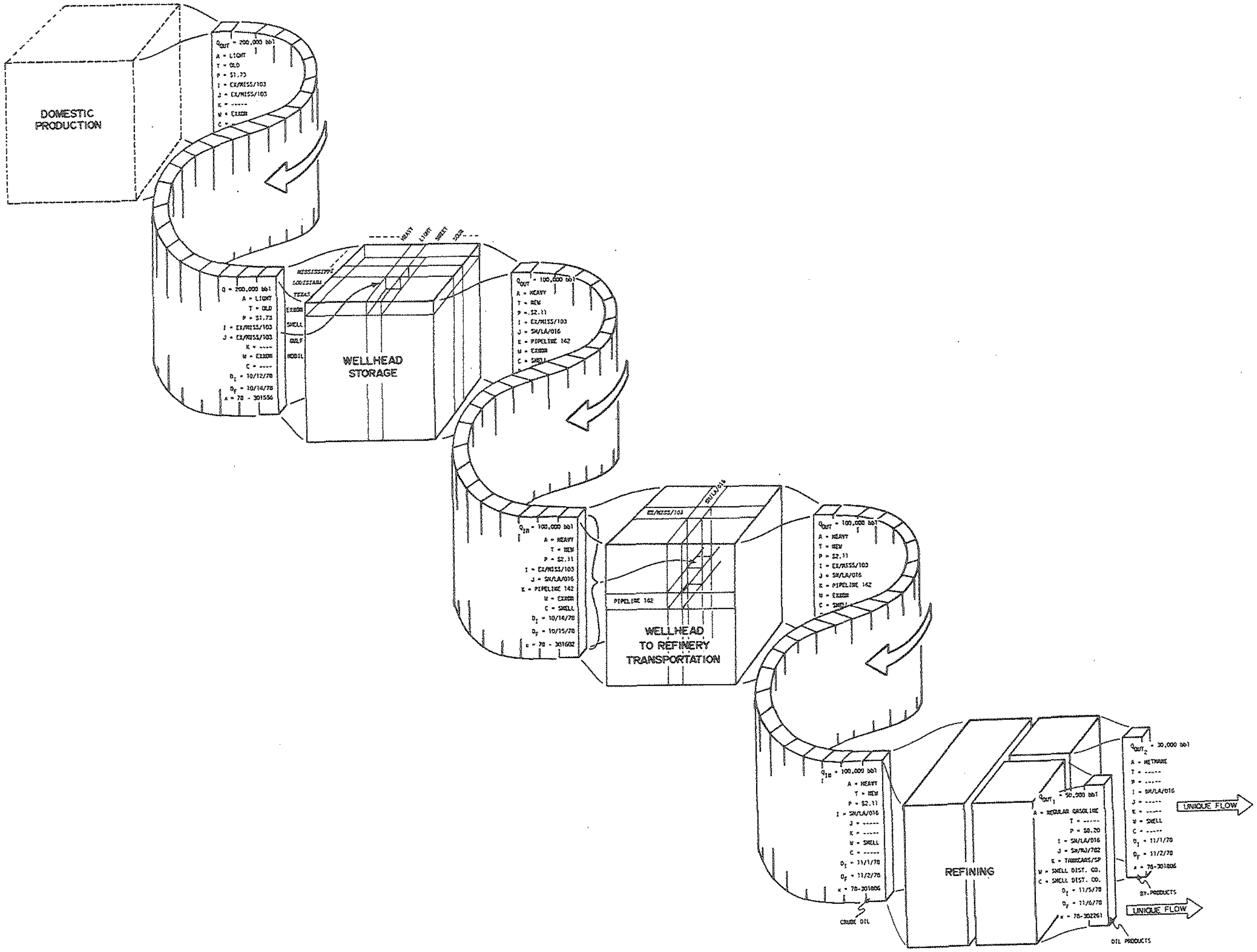


FIGURE 5. Sequential flow through connected blocks representing a portion of the Figure 2 model. Vectorial packets leaving one block flow through a sequential channel restricted to be one unit wide and impinge on the next block downstream. Such incrementation events, properly addressed to specific matrix elements within the blocks, are the only ones allowed to modify the block contents.

XBL 793-9030

A total of five system blocks are depicted in generalized and example form in Figure 3; as described below, they are sufficient to characterize all elements in the system model of Figure 2. In each case, material passes through the block in a unique direction (from acceptance (or inlet) port to release (or outlet) port). At the coupling ports, material packets are represented vectorially - a packet 'shipment' released from a storage element (through its release port) impinges on a transportation element through its acceptance port, and the only contact allowed between the blocks is through the port-to-port interconnection. The ramifications of this channeled flow are three-fold:

(1) Transitions are isolated to the port interface, and a packet is unitized once it has entered a block and cleared the inlet port. This allows storage of single values for each volume Q held within the system blocks, rather than the arrays that would be needed if transitions were allowed to propagate.

Furthermore, a single consignment of crude oil may take several days to pump from, say, a wellhead storage tank into a pipeline that is to take it to a refinery, thus introducing date ambiguity to any attempts to keep track of the incremental flow. Nonetheless, since the consignment is contiguous, it makes sense to treat it as a discrete packet: a consignment of, say, Q = 200,000 barrels, may be transmitted from wellhead storage I to refinery J through pipeline K. This unit treatment has further utility if the wellhead, pipeline and refinery all have different owners; in this case, one must consider not only the unitary sale of Q by the owner of I to the owner of J, but the potential transaction chain of I to K followed separately by K to J. In the last case, the oil actually transmitted by pipeline company K to facility J may not be identically the oil sold by I to K. For simplicity, we wished to concern ourselves at the outset with material flows only, not transactions; however, the transitional identity problem is the same for flows as for transactions, and, in either case, one must consider the pipeline to retain its contents in the same way a storage facility does while the material is entirely contained within it. For transport modes other than pipelines, such as ships or tank-cars, for which loading, movement and unloading are distinct and isolable operations, and for which material is actually contained in tanks while it is being transported, transition isolation appears even more natural.

(2) Incrementation events affecting system blocks may be handled sequentially. Since each vectorial packet, regardless of source or destination (which are variables specified in the attribute vector), passes between blocks in a channel restricted to be only one packet wide, inputs to and outputs from any block in the system affect the block contents one packet at a time. With transitions restricted to the port connections, the internal states of the blocks are subject

to a sequence of controlled modifications and are well defined at any instant. By contrast, if the ports were unrestricted, internal state modification would be continuous and very hard to trace.

(3) Since source (I), destination (J), and other variables in the input and output packet vectors directly address the particular entries in the block matrix that are affected by the packet, and since incrementation events are sequential, the states of the block matrix entries are static except where instantaneously addressed. Thus, stepwise tracing of the effects of particular shipments or other modifications is straightforward; if, for example, some precipitous event happens whose significance is only recognized later. it is possible to recreate the initial system state and step through the subsequent modification events one at a time, thus illustrating, in sequence, the effects of the instantaneous modifications on various parts of the system, and thereby aiding analysis. Such "stop motion" or "instant replay" tracing would be far more difficult (if not impossible) if many modifications were allowed to occur simultaneously as would be the case without the flow restriction.

Similarly, a "sentinel" [10] may be established looking for, say, all flows of gasoline to and from Exxon facilities in New Jersey; establishment of such a sentinel (which could provide, for example, lists of suppliers and customers for Exxon's New Jersey gasoline facilities, together with their volumes added or used, plotted against time) would entail merely accessing the single appropriate point in the refinery product storage matrix and tallying all modifications to it as they happen. This procedural simplicity is an inherent property of the proposed system configuration.

Internally, the blocks themselves are data matrices, as shown by example in Figure 4. For the refinery petroleum product storage block shown in the figure, as for other blocks throughout the system, certain of the transition vector variables address necessarily isolated repositories of product. In all cases, though, the primary variables of interest are the volume (Q) of product on hand and the average price (Pav) paid for that volume. If, for example, an Exxon wholesaler in Metairie, Louisiana, receives 200,000 barrels of kerosene from a Gulf refinery in Galveston, Texas, and puts it in an on-site tank already containing kerosene (which he must do because the products cannot be mixed), then this 200,000 barrels adds selectively to Exxon's kerosene supply in Louisiana. An inlet vector representing the reduced volume impinges on the product storage block and is directed exclusively to cell 211, where it modifies Q_{211} and $P_{av_{211}}$ according to an appropriate set of incrementation equations such as those presented below. (N.B.: As mentioned earlier, the reader is cautioned that the equations presented here, while consistent with the model, are not intended as a final recommendation on appropriate calculation bases. The average price equation, especially, represents a potentially appropriate scheme, but not necessarily an optimal one, and no claims are made for it beyond its use to demonstrate the model being presented here.)

Generalized node modification equations:

$$Q_{\text{final}}(A,I,W,T) = Q_{\text{initial}}(A,I,W,T) + \sum_{\tau} q_{\text{in}}(A,I,W,T) - \sum_{\tau} q_{\text{out}}(A,I,W,T)$$

$$P_{av_{\text{final}}}(A,I,W,T) = \frac{(P_{av_{\text{initial}}} * Q_{\text{initial}}) + \sum_{\tau}(p_{\text{in}} * q_{\text{in}}) - \sum_{\tau}(p_{av_{\text{final}}} * q_{\text{out}})}{Q_{\text{final}}}$$

where τ is a predetermined modification interval (which, in fact, may be set short enough to restrict modification events to occur singly), and outlet flows are considered to occur at the average purchase price rather than actual selling price, so that $P_{av_{\text{final}}}$ will reflect average price paid for the cumulated material at each successive stage.

Sample Application (see Figure 4):

- Initial conditions:

$Q_{211_{\text{initial}}}$(kerosene,EX/LA/103,EXXON,--) = 400,000 bbl.

$P_{av_{211_{\text{initial}}}}$(kerosene,EX/LA/103,EXXON,--) = \$6.57

- Modification effects:

$Q_{211_{\text{final}}}$(kerosene,EX/LA/103,EXXON,--) = 400,000 + 200,000 = 600,000 bbl.

$P_{av_{211_{\text{final}}}}$(kerosene,EX/LA/103,EXXON,--) $= \dfrac{(6.57 * 400{,}000) + (7.26 * 200{,}000)}{600{,}000}$

$= \$6.80$

In the case described, we are assuming that information grain is retained to the level of "corporate x state" totals only; i.e., individual refineries owned by the same company are totaled by state. The block matrix form, though, is universal over all grain levels, and would be just as applicable to the considerably larger matrix that would result if individual refinery information, or even individual tank information, were separately retained. Given an ongoing development process of the data system, this uniformity regardless of grain level is extremely important, since it provides for organized and staged development of what eventually could be an extremely large database.

It should be further noted that aggregation grain affects only the content entries of the blocks, and hence the computer memory requirements of the database. The modification vectors always refer to individual shipment units, i.e., the finest grain structures anticipated, but, since the system takes care of aggregation to whatever level is stored within the blocks through the modification equations, and since only a single modification vector need be in memory at any instant, no penalty is paid in required memory for the fine grain handled between blocks. (Computing time may be another matter, and will have to receive appropriate consideration during system specification.) In terms of reporting, the present burden on respondent companies could be alleviated as well by a data organization as proposed here, since only the raw data need be reported - neither interpretation nor calculation is required on the part of the respondents.

Turning now to the blocks themselves, the five generalized block types naturally form three sub groups: material-conservative, material-nonconservative, and something of a combination. In the ideal case, the storage and transportation blocks are expected to conserve material (although, in practice, losses do occur). Material enters, is routed to an appropriate matrix element and sums with the element's previous contents. In a similar manner, material flowing out of the block is drained only from the appropriate matrix element. Transactions can occur on material without generating its physical movement (and this is a primary reason that we have avoided considering them in the initial model); however, even when the transaction is just an assignment of some volume between branches of a company (without even an exchange of dollars), a modification event has taken place that will be visible to the model (as long as the data is reported). Specifically, a volume (Q) of some type (A), tier (T), etc., of material has changed owner (W); the separate reports for the Q sent from the old owner W_i to the new owner W_j, and received by the new owner W_j from the old owner W_i, should exist just as if a physical shipment had occurred, and would corroborate each other in referencing a specific packet vector. The storage and transportation blocks differ primarily in their dimensionality and in the identity of the variables represented internally within them. In each case, three dimensions are represented in the figures, but the matrices are actually n-dimensional. (While dimensionality higher than three is, of course. hard to display, the computer has no trouble with it.) Input/output vectorial variables that are meaningless for a particular block are ignored.

The second subgroup contains blocks (4) and (5), the sources and sinks. These are not material conservative in that they represent points at which material enters or disappears from the system. In fact, the material does not actually appear or disappear, but, as noted previously in the presentation of Figures 1 and 2, we must define explicit boundaries for the system being modelled that exclude data that we cannot obtain. Thus, for example, we treat crude oil imports as a source element; we cannot be sure of data reported by foreign entities, and consider the source outlet vector at the U.S. border (i.e., U.S. Bureau of Customs import data) as a point at which material simply appears. Similarly, gasoline consumed in, say, the transportation industry, is simply removed from the system at the outlet from the last point to which we trace it (in the present model, the wholesale distributor). As a final example, gaseous refinery by-products leave the oil reporting system and (presumably) enter the natural gas or some other reporting system. Here, we really have a transportation or storage block, but since the outlet port is beyond the bounds of the model system, it does us no good to consider it double-ended, and we treat these by-products as lost, or sinked, instead. Note, however, that here, as elsewhere throughout the model, such losses are explicitly channeled.

The final category of system element is the processing block. In our present model, the sole example of these is the refining block, but it is quite conceivable that further modeling efforts will point to other examples. The present model separates crude oil and refined products into isolated material phases because an acceptably explicit transfer function is not presently available for the refining process. The isolated-phase view necessitates a refinery model in which crude oil is sinked and petroleum products and refining by-products are separately sourced. The sink-source representation emphasizes the fact that our information regarding the refinery is isolated to an accounting of what flows into it and what it produces, and, pending better information as adequate models are developed, still allows us to fruitfully operate the remainder of the system. As in other portions of the model, this allows us to isolate potential error sites from sites at which we expect to have adequate data, and thus to control inaccuracy in the system far better than we could otherwise.

Figure 5 shows what a portion of the Figure 2 flow model might look like when composed out of the generalized blocks that have been presented. The level of monitoring control attained by use of the restricted block interconnections is evident and is, we believe, one of the principal advantages of the present approach. For purposes of illustration, the port interconnections in Figure 5 are represented as vector queues. The connections would be direct, of course, but the intent is to demonstrate that the distinct, but lossless, event sequences at connected ports need not occur in synchrony.

E. Aggregate Model Summary

Summarizing the model as presented, we feel we have developed a viable prototype methodology for the handling of information and data in highly interconnected environments such as the oil energy system treated here. The viewpoint is more important at this stage than the detailed interconnections. Beyond the organizational structure imparted by the representation, the internal matrices of the system blocks directly represent a data structure reasonable to capture the data. In fact, the structure pointed to is an ensemble of databases, connected by a defined set of explicit mathematical relations, and this configuration provides as well the compartmentalization needed to make such a system implementable computationally. While we do not claim to have, at this stage, rigorous completeness or accuracy at the level of the detailed semantics of the oil energy system, we do feel that the methodological structure outlined above contains the essential foundation on which a successful information system can be built in this area of knowledge.

III. APPLICATIONS OF THE MODEL

A. General Query Classes

An information system predicated on the material flow structures outlined here should operate to advantage in responding to several important classes of queries, as indicated below. A major utility of the present approach, in this context, is the direct link between the actual physical flow relationships represented and the matrix storage structure that could reasonably be expected to constitute the eventual data repository; the closeness of this bond between conceptual formulation and data structuring greatly enhances combinatorial data manipulations of all types, including entry, validation, access, retrieval, flagging, and modification.

As has been mentioned, the model as conceived contains no forecasting or prediction capabilities; rather, it is intended as a prototype methodology for organizing information on existing situations, a major undertaking in itself given the complexity of the systems involved. The model is configured, however, to provide the data required in forecasting and prediction, and could be used to support, rather than perform, these functions. A useful definitional bound for forecasting, in this regard, is that of stationarity: a system such as this one can project reasonably only over an interval during which all state variables can be held stationary, i.e., over a single time step, such as, say, one month. Longer term prediction (i.e., what is generally meant by forecasting) cannot be accurate if it is based merely on projection (since projection is not sensitive to cusps and discontinuities in the state variables). Since cusp prediction information is speculative, we consider it best to segregate it from the actual data structure and apply it against the accurate data as an isolated operation. Within the bounds of the stationarity restriction, the data model presented here might answer questions relating to instantaneous remaining emergency capacity, but questions relating to the effects of gasoline price on consumer demand would be, by intent, out of bounds. Similarly, one would not ask this system when the U.S. will run out of crude oil, but the system would be useful, for instance, in suggesting the short term effects of the oil exportation reduction in Iran on New England.

An initial list of query classes addressable by an information system based on the model might contain the following:

(1) Aggregation.

Material or dollar volume totals, or average price, cumulated by state, corporation, facility, usage, industry (e.g., utility or transportation), material, price tier, transport facility, etc., over varying time periods, in any combination (e.g., corporate by state, PAD region, etc.) -- these aggregate data are either identically stored in the system or are immediately calculable from data that are stored. Other data can be either exactly calculated, approximated, or inferred from stored data: profit at each stage, cost versus profit at each stage, value added. Further, cyclical, or otherwise time dependent, information can be selectively accessed and plotted: seasonal importation, consumption, or production (aggregated over U.S., or by state, company, region, etc.), etc.

(2) Process chaining and families.

Supplier/customer communities for individual facilities, regions, companies, products, etc., rank-ordered by product or dollar volume, possibly with seasonal variations. Primary path tracing and distributions for various products from source to end use. (While not addressing supply-demand issues per se, such path tracing might point to driving factors useful in predicting, for example, refinery production.)

(3) Supplies on hand and emergency preparedness.

Volumes instantaneously in system, by product. Instantaneous maximum storage capabilities. Transportation staging and production lag times to respond to various perturbations (such as export cessation from Iran). Regions primarily affected and recovery response time course; re-radiated, reflected and residual effects on various parts of the system of local or focussed perturbations. (Many dynamic system response and wave theory principles have direct application to the state space data here.)

(4) Verification and support data for longer range forecasting, although not the forecasting projections themselves.

(5) Information Validation.

Internal and external validation, cross-checking of data (e.g., monthly facility inventories compared against shipment/receipt data; shipment information from source facility compared against receiving information from destination facility). Data conditional flags and sentinels. (Validation is central to the entire data manipulation process, since the usefulness of the system is compromised unless data accuracy can be assured. Consequently, we treat validation with greater thoroughness in a separate section below.)

(6) System sensitivity to policy modifications.

Given an assumed or estimated consumption efficiency increment, what percentage improvement or regional or U.S. petroleum usage would be expected from a national 55 mph speed limit or from restricting government office buildings to 65° F heating?

Given a new energy extraction process (shale oil, for example) and its associated estimated yields and costs, what percentage of present consumption could it support (and in what regions and product sectors), and how close are present energy costs in these sectors to the point at which the new process becomes cost effective? (Per the discussion introductory to this section, the model addresses present situations only: (1) consumption support of the new process is estimated based on instantaneous present state data only, and is not projected forward; (2) no attempt is made to consider closed-loop effect on the market price of the energy derived, since the present model does not contain a representation for demand.)

B. Data Validation Applications of Model

The model outlined here, while at this stage only skeletal, has a major application in validation and verification of the crude oil and oil products data presently contained in approximately 93 databases and models maintained by the Energy Information Administration. Appendix I is a list of these 93, showing the portions of the model to which they are relevant, and Appendix II provides annotated samples of the information contained in them. The validation benefit of our model lies in the organized view it provides of interrelationships among the data. Given the stated purposes of validation, namely to ascertain the accuracy and relevance of existing data, and to determine in what areas available data is inadequate to give a clear picture of the energy supply chain, this model, or one like it, is an essential tool.

In our view, there are three primary approaches to data validation:

(1) Spot Checks and Audits of Individual Systems.

The data collected by EIA are supplied by respondents on standardized forms originally developed for individual collection systems. One system is examined at a time with a view of streamlining the forms, developing unambiguous definitions of data items and examining the data-collection procedure. In addition, a sample inspection and auditing scheme is instituted in order to ensure that accurate and reliable data is supplied. This is the current approach of the EIA Office of Information Validation. While the approach has the merit of tackling the problem at its roots, it is expensive and relatively cumbersome to apply on a continuing basis to the myriad data collection systems.

(2) Internal Data Validation.

An overview of the data-collection systems as provided in our model shows that data on some important attributes are collected by more than one independent data collection system. In such cases, data collected on a particular attribute may be validated by comparing the data from the independent sources.

For instance, both the Oil Import system (Form FEA-P113-M-0) and the Mandatory Oil Imports Program (Form BOC 7501, 7505) collect data on some of the same attributes of imported oil. These two systems can be used to cross-check each other once the correspondences are identified.

It should be noted that data collection systems used for cross-checking systems must be independent of each other. This is particularly important since some of the data-collection systems are "secondary source systems", i.e., they use data collected by other systems in order to generate their own databases. It is generally possible, then, that two systems derive their data from the same primary data-collection source, and that using them to validate each other will yield meaningless results. In general, one can assume independence if the data forms from which the original data was derived are independent of each other. From a data-validation point of view, it is best if the two forms are not only independent, but have different respondent pools.

Since each data collection system, hence each form, serves a specific purpose, it is rare that any two independent data-collection systems completely overlap each other. One can expect, at best, that one or two of the several attributes on which a form collects data will correspond to attributes on which data is collected by another independent form, and thus constitute usable cross-check sites. Thus, unless a careful study is made of all the individual data-collection forms, it is not possible to identify all internal data validation possibilities. It must be emphasized that particular attention must be paid to the precise definition of the attributes on which data is collected. An attempt is already being made by EIA to standardize definitions and identify similar attributes across various forms through the development of an Information Element Dictionary [5][6].

As previously noted, we have searched the EIA directory of data-collection systems [7] to identify systems which may be collecting data relevant to the petroleum flow model developed in this report. These systems have been grouped in different categories based on the flow model stages and processes to which systems are relevant.

Our preliminary findings, along with a general summary of systems related to the petroleum flow model are presented in Appendix I.

(3) External Data Validation.

This powerful data validation method is available only when all data-collection systems are organized in the form of a composite data-bank, as in our model. This method uses the basic principle of mass or energy conservation. Mass/energy/volume balance equations can be formulated at various stages in the model and appropriate data from independent data-collection systems can be used to verify that the conservation equations are reasonably satisfied; if they are, the data used in the equation are validated. In this technique, the general conservation equation has the following form, and is applicable to volume, mass, or energy balances at individual sites throughout the system:

$$Q_f = Q_i + \sum_{\tau} q_{in}\, dt + \sum_{\tau} q_{out}\, dt$$

where Q_i and Q_f are the initial and final storage volumes/masses/energies at the site, q_{in} and q_{out} are instantaneous flow rates, τ is the accumulation period, and Δt represents, at finest grain, the interval at which data is reported (i.e., the sampling interval) in the systems affecting flows on the branch in question.

Attention should be called to two important aspects of the above equation: (1) the shortest wavelength signal detectable at any site is limited to twice the reporting interval by 'aliasing' [8]; (2) the systems involved are inherently sanpled data systems due to the periodic reporting intervals, and analytical techniques and transforms must be applied in their sampled data domain forms preferentially [9].

The instantaneous form of the conservation equation is more revealing conceptually in its simplicity, and may prove more useful in many situations:

$$\frac{\Delta Q}{\Delta t} = q_{in} - q_{out}$$

A volume conservation equation may be drawn for middle distillates and gives an example of external validation using conservation equations. Here, volume data from the Joint Petroleum Reporting System, the Mandatory Oil Imports System, the Middle Distillate Price Monitoring System and the Bureau of Customs databases may be juxtaposed to validate each other.

Application of appropriate constitutive equations in combination with the conservation equation adds considerably more power to this representation. While these elemental constitutive relations remain to be identified at present, the structure of the model provides the compartmentalization necessary for their explicit definition. Given structural representations of this sort, it is likely that numerous powerful techniques (stability analysis, frequency responses, etc.) from engineering fields such as control theory [8] can find direct application here.

External validation offers the attractive possibility of validating data on an ongoing basis using the computer. However, since we use data from different data-collection systems in the conservation equation, particular attention must be given not only to the compatibility of the attribute definitions, but also to such factors as reporting universe coverage of the data-collection systems, the type of raw-data storage associated with the data-collection system, whether the data-collection system is a primary or a secondary source system, and the general quality and reliability of the individual data-collection systems used in the conservation equation.

Some preliminary work has been done to acquire information on these factors for the data-collection systems related to the petroleum flow model. Wherever possible, individual forms associated with the various data-collection systems have been scrutinized for information on the nature of data attributes addressed by the system, respondent pool characteristics and number, reporting intervals and reporting times, etc. During the course of the present study, many of the relevant forms were not obtainable, and this work is therefore incomplete. Samples of our preliminary findings appear in Appendix II.

Two dynamic characteristics of the reporting systems have dire effects on the comparability of data across systems: reporting frequency (or, inversely, interval) and phase. If a system gathers data quarterly, it can only be compared against others at quarterly intervals. Similarly, if two systems gather data at monthly intervals, but one reports as of the first week of the month and one as of the third, then, especially in a flow situation in which material has continued to move in the intervening interval, any comparisons between them will be indirect; volume anomalies will result due to their phase difference, and the comparison will be proportionally inaccurate. We present a deeper analysis of this synchronization problem in Appendix III, entitled "Data Validation Timing Problems."

IV. CONCLUSION AND SUMMARY

The effort reported here was, as has been stated, preliminary. The intent was to determine how far one could go in applying isolating constraints to a typical energy/economic system, and whether a useful and general model representation could result. From the standpoint of forming a coherent data structure covering such data, this first model has been quite encouraging. The flow structures developed appear general over system aggregation levels and site types, and are amenable to expansion toward greater geographic and temporal flow detail. Further, it appears that the model is appropriate to transaction data as well, and thus we believe it should be developable to adequately represent energy data in the forecasting and regulation domains.

This work was supported by the U. S. Department of Energy, Division of Institutional Relations.

REFERENCES

1. Energy Information Administration Annual Report to Congress, Volume I. (1977)

2. Federal Energy Administration: Project Independence Blueprint, Task Force Report. (PIES Model)
"Analysis of requirements and constraints on the transport of energy materials," Vols. I & II. (November 1974)

3. "World Energy Model: Part I. Concepts and Methods." Energy Modelling: Special Energy Policy Publication. IPC Business Press Ltd. (1974)

4. Federal Energy Administration, "Project Independence Report." (November 1974)

5. Dwyer, B., et al., "Task Force Report on an Information Element Dictionary." Energy Information Administration, DOE, November, 1978.

6. "Energy Validation Information Management." Information Access Corporation, November 1978.

7. Energy Information Administration, Office of Energy Data and Interpretation, "Inventory of data collection system output reports." Washington, D.C., December, 1977.

8. Takahashi, Y., Rabins, M.J., and Auslander, D.M., Control and Dynamic Systems. Addison-Wesley, Reading, Mass., 1970. (Chapter II)

9. Jury, E.I., Sampled Data Control Systems. Wiley, New York, 1958.

10. Rosenberg, S., "SLEUTH, an Intelligent Noticer." Proceedings of the Second National Conference of the Canadian Society for Computational Studies of Intelligence, 1978.

APPENDIX I:

EXISTING DATA COLLECTION SYSTEMS

A review of the 230 existing data collection systems listed in the EIA Directory [7] reveals 63 that are directly related to petroleum and at least 93 that appear peripherally related. Each collects data on one or more of the attribute variables of interest in the model presented in the text. In the table on the following pages, we have listed these data systems in groups corresponding to the model blocks (see text Figure 2) to which they apply, and have indicated the packet and facility vector attributes about which they collect information with stars in the appropriate columns. We have also indexed relevance according to the following schema:

4 = Immediate relevance for model
3 = Subsequent relevance for model
2 = Potential but indirect relevance for model
1 = Background value only
(0 = Not relevant--does not appear in table)

Clearly, information overlaps identified among the data systems imply potential sites for internal data validation. Similarly, contextual and semantic relationships indicated by the model may identify those data systems containing information useful for external validation. Study of these systems in greater depth than is within the present bounds of this project is prerequisite to illumination of such data validation possibilities.

DATA SYSTEMS RELEVANT TO MODEL - SUMMARY

Total data systems (includes models):	230
Related (peripherally or centrally) to our model:	93
Data systems listed under petroleum:	63
Models related to petroleum:	18
Very useful for data validation and model:	29
Useful for data validation and model:	13
Potentially useful for data validation and model:	16
Marginally useful for data validation and model:	9
Peripheral for data validation and model:	7
Forecasting models:	4

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
IMPORTED CRUDE															
1. Foreign Trade Statistical Program	4	*		*		*		*							
2. Fuel Imports by Sulphur Level	3	*		*		*		*							
3. Foreign Crude Oil Cost Report	1	*		*		*		*							
4. Mandatory Oil Imports Project	4	*		*		*		*							
5. Oil Import System	4	*		*		*		*							
6. Petroleum Shipments - Puerto Rico to U.S.	2	*		*		*		*							
7. Transfer Pricing Program	4	*		*		*		*							
DOMESTIC CRUDE															
8. Income and Production Reporting System	2	*		*	*			*							
9. Crude Oil, Natural Gas and Brine Analysis System	2	*		*	*			*							
10. Crude Petroleum Gathered from Leases in Selected States	2	*		*	*			*							
11. District V Monthly Petroleum Report Supplement	2	*		*	*			*							
12. Value at Wells of Crude Oil Purchased	4	*		*	*			*							
13. Land and Mineral Conservation Information Activity	2	*		*	*			*							
14. Crude Oil First Purchaser	4	*		*	*			*							

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
REFINERY INPUTS															
11. District V Monthly Petroleum Report Supplement	2	*		*	*			*	*						*
15. Crude Oil Stocks	4	*		*	*			*	*						*
16. Refinery Operation	4	*		*	*			*	*						*
17. Crude Oil Buy/Sell Program	4	*		*	*			*	*						*
18. Crude Oil Entitlements	3	*		*	*			*	*						*
19. Joint Petroleum Reporting System	4	*		*	*			*	*						*
20. Refinery Cost Pass-Through System	4	*		*	*			*	*						*
REFINING															
16. Refinery Operation	4	*		*				*							*
17. Crude Oil Buy/Sell Program	4	*		*				*							*
19. Joint Petroleum Reporting System	4	*		*				*							*
21. Capacity of Petroleum Refineries	4	*		*				*							*
22. Fuel Consumed for All Purposes at Refineries	4	*		*				*							*
23. Production of Other Finished Products at Petroleum Refineries	4	*		*				*							*
24. Location of Major Oil Refineries	4	*		*				*							*
25. Motor Gasoline System	4	*		*				*							*

DATABASE	RELE-VANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
REFINERY OUTPUTS															
1. Foreign Trade Statistical Program	4	*		*		*	*	*	*						*
16. Refinery Operation	4	*		*		*	*	*	*						*
19. Joint Petroleum Reporting System	4	*		*		*	*	*	*						*
20. Refinery Cost Pass-Through System	4	*		*		*	*	*	*						*
26. Bulk Terminal Stocks of Finished Petroleum Products	4	*		*		*	*	*	*						*
27. Bulk Terminal Stocks of No. 4 and Residual Fuel Oil	4	*		*		*	*	*	*						*
28. Cost and Pricing System	3	*		*		*	*	*	*						*
IMPORTED PETROLEUM PRODUCTS															
4. Mandatory Oil Imports Project	4	*		*		*	*	*							
5. Oil Import System	4	*		*		*	*	*							
6. Petroleum Shipments - Puerto Rico to U.S.	2	*		*		*	*	*							

DATABASE	RELE-VANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
DEMAND - CONSUMPTION															
11. District V Monthly Petroleum Report Supplement	2	*		*		*	*	*	*						*
29. Fuel Oil and Kerosene Sales and Inventories	3	*		*		*	*	*	*						*
30. Sales to Trade and Own Use of Petroleum Products	2	*		*		*	*	*	*						*
31. Market Shares System	4	*		*		*	*	*	*						*
32. Middle Distillate Price Monitoring System	4	*		*		*	*	*	*						*
33. Sup-part L System	4	*		*		*	*	*	*						*
DEMAND - CONSUMPTION CHARACTERISTICS															
28. Cost and Pricing System	3	*		*		*	*	*							
29. Fuel Oil and Kerosene Sales and Inventories	3	*		*		*	*	*							
31. Market Shares System	4	*		*		*	*	*							
34. Fuel Consumption and Cost Statistics	2	*		*		*	*	*							
35. Prices Paid by Farmers for Petroleum Products and Motor Supplies	2	*		*		*	*	*							
36. Annual Survey of Manufacturers	2	*		*		*	*	*							
37. Census of Manufacturers	2	*		*		*	*	*							
see next page															

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
38. Census of Retail and Wholesale Trade	2	*		*		*	*	*							
39. Current Programs of Retail and Wholesale Trade	4	*		*		*	*	*							
40. Monthly Sales of Gasoline by Service Stations	4	*		*		*	*	*							
41. Defense Energy Information System	3	*		*		*	*	*							
42. Sale of Asphalt and Road Oils	2	*		*		*	*	*							
43. Sale of Aviation Fuel	3	*		*		*	*	*							
44. Energy Consumption Database	2	*		*		*	*	*							
45. Compliance Targeting System	4	*		*		*	*	*							
46. Retail Motor Fuels Service Station Survey	2	*		*		*	*	*							
47. Fuel Emergency Report - Oil	1	*		*		*	*	*							

DATABASE	RELE-VANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
TRANSPORTATION: SUPPLY - PROCESSING															
48. Crude Oil and Petroleum Products Pipeline Survey	4	*						*	*	*	*	*	*	*	*
49. Tanker and Barge Shipments of Crude Oil and Petroleum Products in PAD III	4	*						*	*	*	*	*	*	*	*
50. Master Information File Accounting and Financial Systems	4	*						*	*	*	*	*	*	*	*
51. Transportation Movement Information for All Rail Shipments	4	*						*	*	*	*	*	*	*	*
TRANSPORTATION: PROCESSING - CONSUMPTION															
49. Tanker and Barge Shipments of Crude Oil and Petroleum Products in PAD III	4	*		*				*	*	*	*	*	*	*	*
51. Transportation Movement Information for All Rail Shipments	4	*		*				*	*	*	*	*	*	*	*
52. Industrial Energy Conservation Program	4	*		*				*	*	*	*	*	*	*	*
53. Pipeline Movements of Petroleum Products	4	*		*				*	*	*	*	*	*	*	*

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
GENERAL RELEVANCE															
54. International Oil Developments Statistical Survey	1														
55. Annual Survey of Oil and Gas	2														
56. Census Commodity Movement Monthly System	2														
57. Census of Agriculture	2														
58. Financial Accounting System - Naval Petroleum Reserves in California	1														
59. American Petroleum Institute System	2														
60. Natural Gas Liquids Processing Plant Operations	1														
61. Oil and Gas Production System	2														
62. Sales of Liquified Petroleum Gas	2														
63. Regional Energy Policy Project	1														
64. Motor Fuel Consumption by State	2														
65. Coupled Energy System Economic Models	1														
66. Brookhaven Energy Systems Optimization Model	1														
67. Brookhaven Energy Transportation System Submodel	1														
see next page															

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
68. Dynamic Energy System Optimization Model	1														
69. Energy Model Database	1														
70. National Energy Database	1														
71. Crude Oil and Natural Gas Production Model	1														
72. Domestic Crude Oil Pricing Model/System	2														
73. Energy Consumption Database	1														
74. FEA Crude/Transportation Model	1														
75. IEA Quarterly Report	1														
76. International Dynamic Energy Pricing Model	1														
77. International Energy Evaluation System (Model)	1														
78. International Oil Supply Model	1														
79. Neoclassical Regional Growth and Energy Pricing Model	1														
80. OECD Energy Demand Model	1														
81. Oil and Gas Reserves Survey	1														
82. Oil and Gas Supply Model	1														
83. Oil Refinery Yield Model	4														

see next page

DATABASE	RELEVANCE	PACKET-RELATED VARIABLES									FACILITY-RELATED VARIABLES				
		Q	F	A	T	P	U	I	J	K	D	TC	t	CY	C
84. Petroleum Industry Financial Reporting System	1														
85. PIES Integrating Model	1														
86. Refinery and Petrochemical Modeling System	2														
87. Regional Econometric Demand Model	1														
88. Regional Energy Prices and Energy Consumption Model	1														
89. Short Term Petroleum Demand Forecasting Model	1														
90. Short Term Petroleum Supply Model	1														
91. Trends in Refinery Capacity and Utilization	1														
92. Cost and Quality of Fuels for Electric Plants	2														
93. Resource Data Catalog	3														

APPENDIX II

SAMPLES OF DATA COLLECTION SYSTEMS RELEVANT TO AGGREGATED CRUDE OIL AND OIL PRODUCT FLOW MODEL

MODEL STAGE: Imported Crude Oil

SYSTEM: Transfer Pricing Program

FORMS: FEA-F701-M-0

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, A, P

RESPONDENTS: Each refiner which imports at least 500,000 barrels of crude oil a month. The forms are to be filed by the thirtieth calendar day of each month following the month of measurement.

REMARKS: The crude volume is reported by Country Code, Crude Code, API gravity. Also landed cost is reported. Also sulfur content is reported.

1. AGENCY: FEA	2. SYSTEM NAME: Transfer Pricing Program	3. ACRONYM: TPR System	4. STATUS: OPERATIONAL	5. PROCESS MODE: COMPUTERIZED

. PRIMARY SYSTEM USER(S): FEA/RP; FEA/CIA	7. USER SPECIFIC PURPOSE: To monitor and regulate the prices at which oil compares transfer equity crude oil from their foreign to domestic affiliates (Regulatory; Analysis)	8. SYSTEM NUMBER: FEA 018-6047; RP47	9. AGENCY CONTACT: NAME: Doris Dewton FTS TELEPHONE: 254-8660

0. SYSTEMS DESCRIPTION: Petroleum Products

NATURE OF DATA: Information deals with:
1) Imported crude petroleum obtained by purchase and through exchanges,
2) Cost data for imports equity and buy back oil,
3) Crude petroleum sales and purchases, (4) Foreign crude trading activity by country and origin, (5) crude characteristics data.

PROCESSING: Collection and generation

OUTPUT/REPORTS: Hard-copy reports - Data provided information on high, low and average transactio process.

11. RELATED AGENCY PROGRAM: (FEILS)

- USE IN IMMEDIATE PROGRAM: The FEA attempts to control these transfer prices by comparing them with prices from transferring involving the same or similar crude types that were conducted at an arm's-length basis.
- OTHER USERS:

2. FORM NUMBER: FEA-F101-M-0	13. LEGISLATIVE AUTHORITY: P.L. 93-275 10CFR212.84 Implied	14. DATA SOURCES: Government from Non-Government (Oil Companies) PRIMARY	15. VOL./MANDATORY Mandatory	16. CONFIDENT Yes - Input- Public Agency	17. VOLUME: 35 Respond-ents	18. FREQUENC Monthly

19. Methodology: Universe

20. Verification: Audit

MODEL STAGE: Imported Crude Oil

SYSTEM: Petroleum Shipments - Puerto Rico to U.S.

FORMS: FEA-1007-SR

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Limited Q, I.

RESPONDENTS: --

REMARKS: This is a mandatory form filled by all companies that ship crude/unfinished oil/petroleum product to the U.S. Although the data is filed every week, the reporting is by PAD district only. This data can be used, at best, as a validation measure. Also the crude is not categorized by attributes.

NOTE: There are two separate listings for crude oil - to "refiners" and to "other than refiners".

1. AGENCY: FEA	2. SYSTEM NAME: Petroleum Shipments - Puerto Rico to U.S.	3. ACRONYM: None	4. STATUS: OPERATIONAL	5. PROCESS MODE: MANUAL

. PRIMARY SYSTEM USER(S): EIA, Office of Oil & Gas Analysis	7. USER SPECIFIC PURPOSE: Monitor the shipments of crude and products into U.S from Puerto Rico. (Analysis, Policy Formulation)	8. SYSTEM NUMBER: EIA - 2 FEA 030 - 6002	9. AGENCY CONTACT: NAME: Dave Carlton FTS TELEPHONE: 566-9365

0. SYSTEMS DESCRIPTION: Petroleum Products: Imports and Exports

NATURE OF DATA: Data on crude products shipments broken into crude shipments, product shipments, destination (refinery or other), P.A.D. district of entry, custody.

PROCESSING: Tabulated

OUTPUT/REPORTS: Published in M.E.R.

11. RELATED AGENCY PROGRAM: (FEILS)

- USE IN IMMEDIATE PROGRAM: Used to prepare a report on the weekly shipments of crude and produce from Puerto Rico to U.S.

- OTHER USERS: EPAA

2. FORM NUMBER: FEA-1007-SR	13. LEGISLATIVE AUTHORITY: FEA Act PL 93-275 S5,9 EPPA PL 93-159 Implied	14. DATA SOURCES: Sun Commonwealth Gulf Phillips (All Primary)	15. VOL./MANDATORY Mandatory	16. CONFIDENT No	17. VOLUME: 4 Reports 12 Respondents	18. FREQUEN Monthly

19. Methodology: Universe

20. Verification: Telephone

MODEL STAGE: Imported Crude Oil

SYSTEM: Oil Import System

FORMS: FEA-P113-M-0

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, A, I, U

RESPONDENTS: Every company, including subsidiaries or affiliates, which imports crude oil, unfinished oils and finished petroleum products into the U.S. and Puerto Rico. The reporting is monthly and is to be submitted not later than 15 working days after the report month. The reporting period is the calendar month. Report is to be made even if there were no imports during the month.

REMARKS: The import is defined to occur on the "data of withdrawal" from the customs warehouse. For crude oil, the country of origin, port of entry, quantity, sulfur percentage and API gravity are reported. Also reported is whether the oil is for refining/non-refining use and the location of the processing plant.

1. AGENCY:
FEA

2. SYSTEM NAME:
Oil Import System

3. ACRONYM:

4. STATUS:
OPERATIONAL

5. PROCESS MODE:
COMPUTERIZED

. PRIMARY SYSTEM USER(S):
FEA - Office of Regulatory Programs

7. USER SPECIFIC PURPOSE:
To calculate U.S. imports of Petroleum and Petroleum products (Analysis)

8. SYSTEM NUMBER:
FEA 025 - 6253 RP 253

9. AGENCY CONTACT:
NAME: M. Condon
FTS TELEPHONE:

0. SYSTEMS DESCRIPTION: Petroleum Products: Import & Export

NATURE OF DATA: Port of Entry, Country of Origin, Quality of Imports, Import License Numbers, and Product Imported, listed by respondent company.

PROCESSING: Data collection to reports in order that imports be calculated; controlled.

OUTPUT/REPORTS: Monthly Status Report; monthly petroleum statistics report; monthly energy review; quarterly report; quarterly oil and gas report.

11. RELATED AGENCY PROGRAM: FEILS

- USE IN IMMEDIATE PROGRAM:
  Provides a means by which firms report data on the importation of crude oil, unfinished oils, and finished petroleum products into the u.S. and Puerto Rico, as well as shipments of residual fuel oil into the East Coast Refining District.

- OTHER USERS:
  International Energy Agency.

2. FORM NUMBER:
FEA-P113-M-0

13. LEGISLATIVE AUTHORITY:
Presidential Proclamation 3279
P.L. 93-275
P.L. 93-159
IMPLIED

14. DATA SOURCES:
Collected from crude oil and petroleum product importers.
Also:
BOM is another source
PRIMARY

15. VOL./MANDATORY
Mandatory

16. CONFIDENT
No

17. VOLUME:
700 response

18. FREQUE:
12 annually

19. Methodology: Universe

20. Verification: Telephone

MODEL STAGE: Domestic Crude Production

SYSTEM: Crude Oil First Purchaser

FORMS: FEA-P124-M-1

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, T, P

RESPONDENTS: All firms that acquired crude oil through purchase are required to file this form. However, firms with less than 150,000 barrels need file only Schedules A, D and E. Reports have to be filed no later than the first day of the second month following the reporting period which is monthly.

REMARKS: The respondents report volume of all first purchases by 42-gallon barrels separately for Upper Tier (new), Lower Tier (old) and Stripper Well Oils. They also report total volume. However, no reporting is made by oil composition A. Price information is also recorded.

1. AGENCY: FEA	2. SYSTEM NAME: Crude Oil First Purchaser	3. ACRONYM:	4. STATUS: OPERATIONAL	5. PROCESS MODE: COMPUTERIZED

. PRIMARY SYSTEM USER(S): FEA- Regulatory Programs Office	7. USER SPECIFIC PURPOSE: Monitor the prices of domestic crude oil as they apply to existing regulations. (Regulatory)	8. SYSTEM NUMBER: FEA 021-6272; RP 272	9. AGENCY CONTACT: NAME: Xavier Puslowski FTS TELEPHONE: 254-8690

0. SYSTEMS DESCRIPTION: Petroleum Products: Purchases by Type

NATURE OF DATA: Volume and book value of crude, purchases by type (upper tier, lower tier, stripper) by location (state) and by individual producers.

PROCESSING: Calculates the composit monthly price of domestic crude oil and then compares that price with maximum prices permitted.

OUTPUT/REPORTS: Domestic crude oil volume and price analysis summary.
Domestic Crude oil volume and price analysis-company summary.
Purchasers/Sellers report
Volume/costs variance exception report.

11. RELATED AGENCY PROGRAM: FEILS

- USE IN IMMEDIATE PROGRAM:
  The FEA uses the information generated by this system to; 1) measure first sale price of crude, 2) provides information for complicance targeting.

- OTHER USERS:
  EIA, ORP

2. FORM NUMBER: FEA-P124-M-1	13. LEGISLATIVE AUTHORITY: P.L. 94-385 P.L. 93-275 P.L. 94-163 EPCA, ECPA Specific	14. DATA SOURCES: Government collected from any firms that obtained ownership of domestic crude oil through pruchase or other exchange. Primary	15. VOL./MANDATORY Mandatory	16. CONFIDENT Yes Input	17. VOLUME: 250 respon-dents	18. FREQUENCY 12 reports annually

19. METHODOLOGY: N/A

20. VERIFICATION: N/A

MODEL STAGE: Refinery Inputs

SYSTEM: Refinery Cost Pass-Through

FORMS: FEA-P110-M-1

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, P, T

RESPONDENTS: Each refiner, as defined in 10 CFR 212.31. The reporting period is the calendar month,. and the report is to be filed no later than 45 days after the last day of the reporting period.

REMARKS: A fairly complex form involving extensive cost-breakdown computations and data. However, from the point of view of the flow model, the useful reported data is total crude input to refinery in terms of volume and price, refinery fuel usage in volume. Also reported are domestic crude and imported crude bought and the amount of crude resold. Considerable amount of data processing may be necessary to fit into the petroleum flow model.

1. AGENCY: FEA

2. SYSTEM NAME: Refinery Cost Pass-Through

3. ACRONYM:

4. STATUS: OPERATIONAL

5. PROCESS MODE: COMPUTERIZED

. PRIMARY SYSTEM USER(S):
FEA Office of Regulatory Programs

7. USER SPECIFIC PURPOSE:
Used to compute and adjust selling prices of controlled products
(Regulatory; Analysis)

8. SYSTEM NUMBER:
FEA023 - 6008/6105; RP8/105

9. AGENCY CONTACT:
NAME: Andy Drance
FTS TELEPHONE: 254-3426

0. SYSTEMS DESCRIPTION: Petroleum Products: Operations

NATURE OF DATA: Costs and quantities of imported and domestic crude petroleum; Selling prices for covered products

PROCESSING: Collection on a monthly basis

OUTPUT/REPORTS: Monthly hard copy cost summaries for various covered products.

11. RELATED AGENCY PROGRAM: (FEILS)

● USE IN IMMEDIATE PROGRAM: Serves as a means by which refiners subtract to the FEA petroleum pricing regulations, compute and adjust selling prices for covered products (No. 2 oils, jet fuel, gasoline, and propane).

● OTHER USERS:

2. FORM NUMBER: FEA-P110-M-1

13. LEGISLATIVE AUTHORITY:
P.L. 93-275
P.L. 93-159
Implied

14. DATA SOURCES:
Government collects from refiners and natural gas processing plants
Primary

15. VOL./MANDATORY: Mandatory

16. CONFIDENT
Yes
Input
Public
Agency

17. VOLUME: 240 Respondents

18. FREQUE: 12 Reports Annually

19. METHODOLOGY: Universe

VERIFICATION: Audit

MODEL STAGE: Refinery Inputs

SYSTEM: Joint Petroleum Refining System

FORMS: FEA-P302-M-O, FEA-P321-M-O, FEA-P322-M-O, FEA-P323-M-O, BOM: B-01, B-04, B-05, B-09.

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, A ?, I

RESPONDENTS: All trunk pipeline companies which carry crude oil, all refining companies and crude oil producers holding stocks on leases in excess of 1000 barrels in U.S., Puerto Rico and Virgin Islands. Data is to be reported on crude oil stocks as of midnight of the last day of the reporting month. No reporting deadline is indicated on form FEA-P323-M-O.

REMARKS: Only form FEA-P323-M-O carries crude oil information. FEA-P302-M-O and all the BOM forms were not available at the time of preparation of this report. The form FEA-P323-M-O breaks down crude oil stocks statewise, as domestic or foreign. Also, stocks are broken down into Refinery stocks and Pipeline and tankfarm stocks. Also state of origin is given separately so that we can get some idea of type of crude, A, for domestic crude.

1. AGENCY: FEA	2. SYSTEM NAME: Joint (FEA/BOM) Petroleum Reporting System	3. ACRONYM: JPRS	4. STATUS: OPERATIONAL	5. PROCESS MODE: COMPUTERIZED

. PRIMARY SYSTEM USER(S): FEA/BOM	7. USER SPECIFIC PURPOSE: To monitor production and stocks of petroleum crude and refined products (Regulatory)	8. SYSTEM NUMBER: FEA 005 (6230/6301) EIA 230	9. AGENCY CONTACT: NAME: Pat Holmes FTS TELEPHONE: 254-8450

0. SYSTEMS DESCRIPTION: Petroleum Products

NATURE OF DATA: Refinery Production; receitps, inputs, shipments or losses. Finished petroleum stocks, imported foreign crude oil.

PROCESSING: FEA requests data from BOM, which collects and edits data from (PAD); and outputs reports monthly

OUTPUT/REPORTS: Monthly energy review
Monthly petroleum statistics report

11. RELATED AGENCY PROGRAM: (FEILS)

- USE IN IMMEDIATE PROGRAM: The system combined the petroleum reporting requirements of FEA/BOM as well as the Department of the Interior

- OTHER USERS: Department of Interior
Petroleum Administration for Defense (PAD)

2. FORM NUMBER:	13. LEGISLATIVE AUTHORITY:	14. DATA SOURCES:	15. VOL./MANDATORY	16. CONFIDENT	17. VOLUME:	18. FREQUENC
FEA-P302-M-0 FEA-P321-M-0 FEA-P322-M-0 FEA-P323-M-0 BOM: B-01 B-04 B-05 B-09	FEA Act of 1974 P.L. 93-275; G.O. 11790 & 39 CFR 23185 EPAA of 1973 (P.L. 93-159) Specific	FEA, Office of Data (Primary) BOM/DOI Petroleum bulk terminal operators, operators of pipelines, and all refineries (Secondary)	Mandatory	Yes Input Public	255 Reports	12 Annually

19. METHODOLOGY: Universe

20. VERIFICATION: Check for inconsistencies

MODEL STAGE: Refinery Inputs

SYSTEM: Crude Oil Entitlements

FORMS: FEA-P102-M-1, FEA-P103-M-O, FEA-P126-M-O, FEA-P129-M-O.

VARIABLE(S) FOR WHICH DATA IS COLLECTED:

Q, T

RESPONDENTS: FEA-P102-M-1, FEA-P103-M-O are both filed by all refiners of crude oil. The reporting period is a calendar month, and the report is to be filed by 5th day of the second month following the reporting month for FEA-P102-M-1 and on the 10th day of the reporting month for FEA-P103-M-O. The periods and filing deadline for FEA-P126-M-O are the same as FEA-P102-M-1. This form is to be filed by all importers of residual fuel oil.

REMARKS: The useful information is in terms of runs to stills and oil receipts which contain price-tier information.

1. AGENCY: FEA

2. SYSTEM NAME: Crude Oil Entitlements

3. ACRONYM: FEA-90

4. STATUS: OPERATIONAL

5. PROCESS MODE: COMPUTERIZED

. PRIMARY SYSTEM USER(S): FEA Regulatory Programs Entitlements Program

7. USER SPECIFIC PURPOSE: To establish the monthly entitlements buy/sell position of each participant
REGULATORY: ANALYSIS

8. SYSTEM NUMBER: FEA019-6072; EIA5

9. AGENCY CONTACT:
NAME: Doris Dewton
FTS TELEPHONE: 254-8660

0. SYSTEMS DESCRIPTION: Petroleum Products: Sales/Purchase Entitlements

NATURE OF DATA: Weighted average costs by various crude categories, adjustments to estimated volumes for crude, total curde runs to stills, required purchase/sale of entitlements, bias and exception relief, and domestic crude oil supply ratio.

PROCESSING: Collection and process

OUTPUT/REPORTS: Federal Energy Guidelines, Regulatory Management Report; Historical Cost Comparison; Calculations Report; Federal Register Report, Processing Agreement Cross-Check.

11. RELATED AGENCY PROGRAM: (FEILS)

- USE IN IMMEDIATE PROGRAM: To collect and process data on crude oil purchases and support the crude oil allocation program for this purpose of insuring the maintenance of competitive domestic market place for all refiners.

- OTHER USERS:

2. FORM NUMBER:
FEA-P102-M-1
FEA-P103-M-0
FEA-P126-M-0
FEA-P129-M-0

13. LEGISLATIVE AUTHORITY:
P.L. 93-159
10 CFR 211.66 (h)
EPAA

IMPLIES

14. DATA SOURCES: Government from importers and refiners

PRIMARY

15. VOL./MANDATORY: Mandatory

16. CONFIDENT: Yes

INPUT
Public
Agency

17. VOLUME: 200

18. FREQUE: Monthly

19. METHODOLOGY: Universe

VERIFICATION: Audit

APPENDIX III

DATA VALIDATION TIMING PROBLEMS

Synchrony of data reporting is a prerequisite of interval-based data collection systems. Significant over- or under-counting errors can be introduced into such systems by variation in the instant of cumulation (both within and across respondent populations), and since these errors are compounded whenever systems with different reporting intervals or cumulation dates are combined, validation of data in these systems is not a simple task. In the following pages, we present an analysis of the error bands associated with data validation in synchronous and asynchronous systems.

The problems outlined here support our suggestion that the aggregated petroleum model be based on data from shipment packets themselves rather than on monthly or quarterly cumulations. In the case of data validation involving asynchronously reporting data collection systems, one can estimate statistically the volume of material expected to be in transition at any instant; this transition volume can generally be expected to represent a reasonably small and constant proportion of the total. On the other hand, if it were possible to organize all the data-collection systems synchronously, the timing problems associated with data validation would be greatly reduced.

Various data collection systems monitor a set of petroleum volumes $x_1, x_2, \ldots, x_n$, which may then be organized into a material conservation (data validation) equation of the form,

$$x_1 \pm x_2 \pm \ldots \pm x_n = 0$$

1) Each of these data is collected at different intervals depending upon the data collection system. Let $T_1, \ldots, T_n$ represent the sampling intervals corresponding to $x_1, \ldots, x_n$ respectively.

2) We shall assume that starting from some point in time these data are collected at regular intervals as specified above. Let t_{01} represent the starting point of data x_1 so that x_1 is first reported at time t_{01}.

3) The data can be expressed in time series as follows:

$$x_1\ (t_{01}),\ x_1\ (t_{01} + T_1),\ \ldots,\ x_1\ (t_{01} + nT_1),\ \ldots$$

$$x_2\ (t_{02}),\ x_2\ (t_{02} + T_2),\ \ldots,\ x_2\ (t_{02} + nT_2),\ \ldots$$

$$x_n\ (t_{0n}),\ x_n\ (t_{0n} + T_n),\ \ldots,\ x_n\ (t_{0n} + nT_n),\ \ldots$$

4) Some of the x_i are flow volume data while others are storage volume data. Since these have to be analyzed differently later, let us distinguish between these by denoting flow variables by x_i and storage variables by s_j. Under this new notation, a data-validation equation can be represented in a general way as follows:

$$x_1 \pm x_2 \ldots \pm x_n + \Delta\ (\pm\ s_1, \pm\ s_2 \ldots, \pm\ s_n) = 0$$

where Δ denotes the change in the value of the variable from its previous value.

Using the following further notations,

Sampling intervals for the x_i : T_i

Sampling intervals for the s_j : U_j

Starting time for the x_i : t_i

Starting time for the s_j : u_j ,

the data are now in the form

$$x_1\ (t_1),\ x_1\ (t_1 + T_1),\ \ldots,\ x_1\ (t_1 + nT_1),\ \ldots$$

$$x_i\ (t_i),\ x_i\ (t_i + T_i),\ \ldots,\ x_i\ (t_i + nT_i),\ \ldots$$

$$x_n\ (t_n),\ x_n\ (t_n + T_n),\ \ldots,\ x_n\ (t_n + nT_n),\ \ldots$$

$$S_1\ (u_1),\ S_1\ (u_1 + U_1),\ \ldots,\ S_1\ (u_1 + nU_1),\ \ldots$$

$$S_m\ (u_m),\ S_m\ (u_m + U_m),\ \ldots,\ S_m\ (u_m + nU_m),\ \ldots$$

The x_i's represent flow volume during the reporting interval while the S_j's represent the storage volumes at the end of the reporting interval. i.e., $x_i\ (t_i + T_i)$ represents the volume of flow in the interval $(t_i,\ t_i + T_i)$ while $S_j\ (U_j + U_j)$ represents the storage volume at the instant $(U_j + U_j)$.

The timing problems can be examined as follows:

Case A. Synchronized Data Collection System:

Case 1 This implies that $t_1 = \ldots = t_n = u_1 = \ldots = u_m$ i.e., all the data collection systems start at the same point.

Also $\{T_i\}$, $\{U_i\}$ can be arranged in an ascending order such that each element of the series in an integer multiple of the previous element in the series.

In this case the shortest data validation interval, DT, will be determined by the longest sampling time of among all all the variables in the equation.

i.e., $DT = \text{Max}\ [\{T_i\}, \{U_j\}]$

If the data is validated at times $d_1, d_2, \ldots, d_s$ such that $d_{k+1} - d_k = DT$, the validation equation will take on the following form when carried out at time d_{k+1}.

$$\left\{ \sum_{i=1}^{n} \sum_{\ell=1}^{\left(\frac{DT}{T_i}\right)} \pm x_i \ (d_k + \ell T_i) \right\} + \left\{ \sum_{j=1}^{m} \pm \ [S_j \ (d_{k+1}) - S_j \ (d_k)] \right\} = 0$$

Note: Analytically this is quite rigorous, except that we may end up with situations where DT is quite large; i.e., we may not be able to validate data as frequently as we like.

Sub Case In addition to synchronization, we also have

$T_1 = T_2 = \ldots = T_n = U_1 = \ldots = U_m = T$

i.e., we ensure that all data is gathered at the same sampling instants and with same sampling intervals. This, of course, is the ideal case.

The data validation equation then takes on the following simple form:

validation interval DT = T

for the k^{th} validation, the equation is

$$\sum_{i=1}^{n} \pm x_i \ (kT_i) + \sum_{j=1}^{m} \pm [S_j \ (kT) - S_j \ ((k-1)T)] = 0$$

Case 2 Synchronized starting but with non-synchronous sampling intervals.

i.e., $t_1 = t_2 = \ldots = t_n = u_1 = \ldots = u_m$

but $T_1 \neq T_j$ or $U_i \neq U_j$ at least for some i, j

In this case, the shortest possible sampling interval is the least common multiple of all T_i, U_j

Thus, $DT = \text{LCM} \left\{ \{T_i\}, \ \{U_j\} \right\}$

where LCM stands for Least Common Multiple. For the $(k+1)^{th}$ validation, (which will occur at t = [DT x (k+1)]), and the validation equation has a similar form as before.

$$\left\{ \sum_{i=1}^{n} \sum_{\ell=1}^{\left(\frac{DT}{T_i}\right)} \pm x_i \ (d_k + \ell T_i) \right\} + \left\{ \sum_{j=i}^{m} \pm [S_j \ (d_{k+1}) - S_j \ (d_k)] \right\} = 0$$

where $d_k = (DT \times k)$

Sub Case

$t_1 = \ldots = t_n = u_1 = \ldots = u_n$

and $T_i \neq T_j$ and/or $U_i \neq U_j$

but all T_i, $U_j = K \text{ Min } [\{T_i\}, \ \{U_j\}]$

where K = integer.

This is a simpler version of above, and the solution will have the same form as above.

The only difference is that

$DT = LCM \left\{\{T_i\}, \{U_j\}\right\}$ will be such

that $DT = K \text{ Min } [\{T_i\}, \{U_j\}]$

where $K = \text{Max } [\{T_i\}, \{U_j\}] \Big/ \text{Min } [\{T_i\}, \{U_j\}]$

Case B. Asynchronously Started Data Collection

This implies that the data collection is not started at the same time. The sampling intervals may or may not be different, but the starting times are different.

i.e., $t_i \neq t_j \text{ V } u_i \neq u_j$ at least for some i, j

Note: V stands for and/or

Case 1 Asynchronous starting with unequal sampling intervals. (This is the simpler case.)

$t_i \neq t_j \text{ V } u_i \neq u_j$ for some i, j

and $T_k \neq T_\ell \text{ V } U_k \neq U_\ell$ for some k, ℓ

In such a case as this, the first data validation point (where all data is recorded at the same time), is given by

$t = d_1$ such that

$$d_1 = (t_1 + K_1T_1) = (t_2 + K_2T_2) = \ldots = (t_n + K_nT_n) = (u_1 + L_1U_1)$$
$$= (u_2 + L_2U_2) = \ldots = (u_m + L_mU_m)$$

where all $K_1, \ldots, K_n, L_1, \ldots, L_m$ are integers.

From this point on, the validation intervals are given by DT such that

$$DT = LCM \left\{\{T_i\}, \{U_j\}\right\}$$

For the first data validation, at $t = d_1$ the data validation equation is

$$\sum_{i=1}^{n} \sum_{k=1}^{\left(\frac{d_1-t_1}{T_i}\right)} \pm x_i\ (t_i + kT_i) + \sum_{j=1}^{m} \pm [S_j\ (d_1) - S_j\ (u_j)] = 0$$

Note: We have to assume that S_j (o) ≃ S_j (u_j). Otherwise we are in trouble because S_j (o) is not recorded.

For the $(k+1)^{th}$ data validation point, the equation is

$$\sum_{i=1}^{n} \sum_{\ell=1}^{\left(\frac{DT}{T_i}\right)} \pm x_i\ (d_k + \ell T_i) + \sum_{j=1}^{m} \pm [S_j\ (d_{k+1}) - S_j\ (d_k)] = 0$$

Note: It is perfectly possible that we could end up with an unacceptably large value for d_1. However, this is dictated solely by our insistence on rigor.

Case 2 Asynchronous starting times with equal sampling intervals.

$t_i \neq t_j \vee u_i \neq u_j$ for some i, j

and $T_1 = \ldots = T_n = U_1 = \ldots = U_m$

Sub Case 1

$t_i \neq t_j \vee u_i \neq u_j$ for some i, j

and $T_1 = \ldots = T_n = U_1 = \ldots = U_m = T$

and $t_i = t_j + KT_j$ where K = integer

and $u_i = u_j + KU_j$

This is a simple case. This means that the starting points are not the same, but at the longest starting times all samples are available.

The first data validation point d_1 is

$$d_1 = \max \ [\{t_i\}, \{u_j\}]$$

$$DT = T$$

and the validation equations are as follows at $t = d_1$

$$\sum_{i=1}^{n} \sum_{\ell=0}^{\left(\frac{d_1 - t_1}{T_i}\right)} \pm x_i \ (t_i + \ell T_i) + \sum_{j=1}^{m} \pm [S_j \ (d_1) - S_j \ (u_j)] = 0$$

for the $(k+1)^{th}$ validation which takes place at

$$\begin{aligned} d_{k+1} &= d_1 + K \ DT \\ &= d_1 + KT, \end{aligned}$$

the validation equation is

$$\sum_{i=1}^{n} \pm x_i \ (d_k + T) + \sum_{j=1}^{m} \pm [S_j \ (d_k + T) - S_j \ (d_k)] = 0$$

where $d_k = d_1 + (k-1)T$

<u>Sub Case 2</u>

$$t_i \neq t_j \ V \ u_i \neq u_j \text{ for some } i, j$$

and $T_1 = \ldots = T_n = U_1 = \ldots = U_m$

and $t_i \neq t_j + KT_j$ where K = integer

$V \ u_i \neq u_j + KU_j$ for some i, j

This is by far the most difficult case.

It is easy to show that there is no t = d such that

$d = t_1 + KT_1 = \ldots = t_n + K_nT_n = u_1 + L_1U_1 = \ldots = u_m + L_mU_m$

where $K_1, \ldots, K_n, L_1, \ldots, L_m$ are integers.

Note: This means that there is no instant in time where all data are sampled simultaneously.

The best procedure in this case is to find a time at which we have as many samples as possible. At this instant we estimate the unavailable data using some smoothing technique.

Note: If t = d is the point in time at which all data except x_i are sampled, we could estimate x_i (d) using the immediately previous and immediately next samples.

i.e., use x_i (<d) and x_i (>d) to estimate x_i (d)

we can very simply estimate

$$x\ (d) = \frac{d-t_p}{T} [x_i\ (t_n) - x_i\ (t_p)] + x_i\ (t_p)$$

where x_i (t_p) is the measurement at $t = t_p$ immediately prior to t = d

x_i (t_n) is the measurement at $t = t_n$ immediately after t = d

t is the reporting interval for data x_i

If we need more refinement, we can use a Langrangian polynomial interpolation, provided we have a sufficient number of previous data points.

Summary

The data-validation timing problems considered in this section pertain specifically to the time interval between successive data-validations. In most of these cases, we have determined the time interval under the constraint that all data required by the data-validation equation be available at the same time. It is of course possible to use some statistical procedure to estimate the unavailable data at any arbitrary time, but we have not seriously considered procedures of this type in this analysis. The results obtained in the preceding pages can be briefly summarized as follows:

1. In the ideal validation case, all data collection systems are started synchronously and have the same reporting interval T. In such a case, a rigorous data validation is possible at every sampling time and the data-validation interval is the same as the reporting interval.
2. A less desirable case occurs when the data collection systems are all started synchronously but have different reporting intervals. In such a case, rigorous data validation is possible only at very large time intervals, and the duration of the interval between data validation instants is given by the least common multiple of the individual sampling intervals of the various data collection systems involved in the data validation equation. This, of course, can result in unacceptably large time periods between successive data validations.

   Although the above case is not common in actual practice, a special subcase of the situation is very common in existing data collection systems. In this case, the starting points are synchronized but the data reporting intervals of the various systems are unequal. However, the intervals are such that they are all integer multiples of the smallest reporting interval. Such a case would exist when a data validation equation involves systems whose reporting intervals are monthly, bi-monthly, quarterly, half-yearly, etc. In this case the effective data-validation interval is the largest reporting interval of the data-collection systems represented in the data-validation equation.

3. The case of asynchronously started systems is somewhat more complicated. When asynchronously started data-collection systems also have unequal reporting intervals, rigorous data validation is possible only at large intervals given by the least common multiple of the reporting intervals of the systems involved. An added disadvantage is that the first data validation point might itself take a very long time after the starting time.

A special and rare subcase of the above occurs when the asynchronously started systems have equal sampling times. If the starting times are completely random (which, fortunately, is not the case) then it may not be possible to obtain a rigorous data validation. One has no options but to resort to some sort of an estimation procedure in this case.

In final summation, it should be noted that the entire timing asynchrony question is avoided in validation systems based on instantaneous incrementation rather than cumulative reporting, a decided advantage given the practical difficulty of coordinating such systems in the real world.

U.S.GPO:1979-689-058(F)30

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720