This paper is the second step in the preparation of forecasts of occupational and industrial information which will meet the needs of the Information System for Vocational Decisions (ISVD). The author discusses the computation routines which need to be developed, tested and operationalized toward the goal of combining occupational and industrial information and projections, and storing, processing and retrieving it. The paper begins with a fairly abstract discussion of the terminology and principles to be used. These principles are then applied to the collection of information from the available sources. (TL)
INFORMATION SYSTEM FOR VOCATIONAL DECISIONS

Project Report No. 6

FORECASTING FOR COMPUTER-AIDED CAREER DECISIONS:
PROSPECTS AND PROCEDURES

Richard M. Durstine

This paper was supported in part by Grant No. OEG-1-6-061819-2240 of the United States Office of Education under terms of the Vocational Education Act of 1963.

Graduate School of Education
Harvard University

March, 1967
FORECASTING FOR COMPUTER AIDED CAREER DECISIONS:
PROSPECTS AND PROCEDURES

Richard M. Durstine
Graduate School of Education
Harvard University

Introduction

This paper is the second step in the preparation of forecasts of occupational and industrial information. It extends and develops the ideas of Russell G. Davis' recent survey of forecasting methodology (Reference 1), using forms, procedures and work programs designed to the needs of the ISVD project.

The eventual aim, for which these necessary foundations are now being laid, is to combine information from diverse sources and thus to provide forecasts more complete and comprehensive than are now available. This can be done only through carefully designed computation routines, which will take some time to develop, test and put into operation. The present paper fills the gap between the basic methodology (Reference 1) and the working routines. It should also serve as a basis for discussion in planning for and preparing those routines.

The following are the goals of this effort:

a) Ability to collect and absorb in explicit form and with minimum distortion any objective statement about the future of occupations or industries and their attributes,
and to relate such statements to one another.

b) Specifically, projections of employment, earnings, etc., for future years by occupation and by industry, in as much detail as the available sources of information permit.

c) Separate treatment of short and long range projections, making use of different sources of information for these two classes of projections.

d) Provision for

Finding
Collecting
Organizing
Storing
Processing
Retrieving

information in as general a form as possible.

The underlying attitude throughout is that the limited resources of ISVD make it presumptuous to prepare new forecasts, except on an experimental basis. The primary job will therefore be to assemble and integrate what has been prepared by others, to interpolate where gaps exist, and to identify deficiencies. Experimental development of further forecasting capability is possible, but only after existing material has been thoroughly tapped.

The remaining discussion will begin with a fairly abstract discussion of the terminology and principles to be used. This will be followed by the application of these principles to the collection of
information from the available sources. This approach will then be shown to be a consistent extension of the methods given in (1).

Method and Terminology

The logical constructs and terminology that will be used repeatedly in developing forecasts for career decisions will now be explained briefly. If the reader finds this description excessively abstract, he can proceed directly to the next section, where the presentation is more applied and explicit, and use this earlier material as a reference.

Information will be identified here in terms of coordinate dimensions and content dimensions, where

a) **Coordinate dimensions** describe the situation (e.g., industry, occupation, time) to which the information refers. They can be thought of as the identifying labels on the rows and columns of a table or matrix.

b) **Content dimensions** describe the nature of the information (e.g., population, average earnings, level of employment). Clearly, what is a coordinate dimension in one instance may be a content dimension in another, so the two are not always distinct. In context, however, they will usually be distinguishable from one another.

The dimensions (particularly the coordinate dimensions) can be separated into

a) **Scaled dimensions**, to which a numerical scale, either continuous or discrete, can be attached. These dimensions
admit both to values specific to a particular point on
the scale, and to averages or totals over intervals of
the scale. *Time* and *age* are such dimensions.

b) **Unscaled dimensions**, to which no scale can be attached. These
dimensions must be broken into exhaustive and mutually exclusive
categories. Examples are *industry* and *occupation*.

When dimensions are unscaled or when scaled dimensions are treated in
terms of intervals, the total range of possibilities covered will be
called the **domain** of the dimension, and the exhaustive and mutually ex-
clusive set of categories or intervals that cover the domain will be
called its **partition**. A domain can have several distinct partitions,
of course.

**Information content**, when expressed in quantitative terms, can be
given as:

a) **Total quantity** associated with a relevant point, interval,
or category.

b) **Level** (e.g., average value) of the quantity within a category
or interval. This level will relate to **total quantity** through
some measure on the category or interval. (e.g., *Wage level*
for an occupation is related to **total wages** through the number
of persons pursuing that occupation. Here number of persons
is the measure, and the individual occupations are the
categories.)

c) **Fraction** of the total quantity in the domain that is contained
by a category or interval.

For theoretical work and for abbreviated identification, the fol-
lowing nomenclature will be used to describe information content:
where the subscript \( j \) refers to the type of content (population, earnings, etc.) which must always be clearly identified. Clearly, change and rate of change must always be expressed in terms of some scaled coordinate dimension.

Let the domain of a dimension (or of a space spanned by several dimensions) be represented by \( \Delta_m \) and the partitions within \( \Delta_m \) be represented by \( P_m \).

Let the individual cells of the partition be identified by the index \( h \). Then

\[
\sum_h q_j(h) = 1 \quad \text{(fractional parts must sum to the whole)}
\]

\[
\sum_h s_j(h) = \sum_h r_j(h) = 0 \quad \text{(changes in fractional parts must cancel out)}
\]

where \( q_j(h), Q_j(h), \text{etc.} \) will be used as short forms, where

\[
q_j(h) = q_j(P(h)) = q_j(P_{mn}(h))
\]

\[
Q_j(h) = Q_j(P(h)) = q_j(P_{mn}(h))
\]

etc.

The shortest form consistent with clarity will usually be used. Also

\[
Q_j(h) = W_1(h) Q_j(h)
\]
where $W_i(h)$ is the weighting function or measure, mentioned earlier, defined on $P_{mn}$, that relates level to total quantity. The index $i$ is included to distinguish among such measures. Note that $W_i$ is itself a form of information content, renamed to emphasize the special purpose it serves here.

Another relevant general formula is

$$q_j(h) = \frac{Q_i(h)}{\sum_h Q_j(h)}$$

which converts totals to fractional parts that sum to unity over the domain $\Delta_m$.

The total of $Q_j$ over the domain $\Delta_{mn}$ will be denoted by $Q_j(\Delta_{mn})$, where the subscript $n$ is needed in case there is a difference in $Q_j$ depending on the partition of $\Delta_m$. Then

$$Q_j(\Delta_{mn}) = \sum_h W_i(h) \bar{Q}_j(h)$$

and the level of $Q_j$ over the entire domain is

$$Q_j(\Delta_{mn}) = \frac{\sum_h W_i(h) \bar{Q}_j(h)}{\sum_h W_i(h)}$$

So $\sum_h W_i(h)$ plays the same role in relating $\bar{Q}_j(\Delta_{mn})$ to $Q_j(\Delta_{mn})$ as did $W_i(h)$ between $\bar{Q}_j(h)$ and $Q(h)$. 
Collection and Preparation of Information

For applied work, domains and partitions must be identified in terms of specific individual coordinate dimensions, from which more complex domains and partitions can be constructed as needed.

Information will be gathered from a variety of sources, stored in a consistent manner, and combined to be used, insofar as possible, as an integrated whole. Some of the more fruitful of these sources are shown in References 2 through 4.

The procedure to be followed in gathering, preparing, and treating this data will be as follows:

a) Survey available information sources. To this end, procedures for both preliminary and detailed surveys must be devised. These procedures will be outlined later in this memorandum.

b) Collect and store this information.

c) Devise routines for its manipulation, and in particular for improvising information that is not directly available from the original sources.

The intent is to have a structure able to treat a broad range of information. We thus need a knowledge of which information is, or is likely soon to be, available. Experimentation with small segments of this information will serve to test out the structure. Subsequent collection, inclusion, and use of information will depend on its availability and on the needs of the ISVD project. The goal is a working tool that can then be used and progressively developed. We seek a living, growing organism, not a closed, static data base.
In this sense the job can never be finished, but our results should be all the more valuable because of this trait.

The following paragraphs set forth a program for developing this information gathering capability, with example procedures.

**Tentative Work Schedule**

A comprehensive survey of sources of information and their subsequent development into the proposed system for information treatment will involve several steps. These are listed below in terms of approximate sequence, and type of personnel who would be the principal participants.

<table>
<thead>
<tr>
<th>Professional Personnel</th>
<th>Clerical Personnel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preliminary survey of information sources</td>
<td>2. Full survey of information sources</td>
</tr>
<tr>
<td>3. Preparation of formulas</td>
<td>5. Ongoing assembly, punching, and verification of data</td>
</tr>
<tr>
<td>4. Preparation of computation routines</td>
<td></td>
</tr>
<tr>
<td>6. Ongoing experimentation and development</td>
<td></td>
</tr>
</tbody>
</table>

**Procedures for Collection of Information**

The preliminary and full surveys of information sources will be described here in terms of the forms to be used for these surveys and in terms of example dimensions, domains, and partitions.

Forms for the preliminary information survey are as follows:

1. **Source List** (see Exhibit 1)
2. **Catalog of Content Types** (see Exhibit 2)
3. **Simplified Catalog of Domains** (see Exhibit 3)

4. **Simplified Catalog of Partitions** (see Exhibit 4)

5. **Preliminary Source Survey** (see Exhibit 5)

<table>
<thead>
<tr>
<th>Source Number</th>
<th>Title</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>America's Industrial and Occupational Manpower Requirements</strong>, etc.</td>
<td></td>
</tr>
</tbody>
</table>

Exhibit 1

Example Form for **Source List** (for both preliminary and full surveys)

In the **Source List** of Exhibit 1, each source of information is assigned a number, identified by title, etc., and the location of a copy of the source is indicated. In the example shown above, a library number related to the project collection is used.
### Form of Contents

<table>
<thead>
<tr>
<th>Serial Number (Content type)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>Population (persons)</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Population (households)</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>Proportion of population (persons)</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Annual rate of change of family income</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>Proportion of total income</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Exhibit 2

**Example Form for Catalog of Content Categories**

(for both preliminary and full surveys)

A **Catalog of Content Categories** (Exhibit 2) is needed to keep track of the quantities that are being included as content dimensions, to insure consistency of notation and to avoid repetition. Designation of the form of the contents here is consistent with the scheme suggested earlier in this paper (i.e., Q for total quantities, q for fractions of the whole, etc.). The serial number specifies to what type of information the content refers. There need be no system to the assignment of serial numbers, since they are used for identification only.
<table>
<thead>
<tr>
<th>Domain Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>All ages</td>
</tr>
<tr>
<td>L1</td>
<td>All locations including armed forces overseas</td>
</tr>
<tr>
<td>L2</td>
<td>All locations, not including armed forces overseas</td>
</tr>
<tr>
<td>I1</td>
<td>All industries</td>
</tr>
<tr>
<td>I2</td>
<td>All manufacturing industries</td>
</tr>
</tbody>
</table>

Exhibit 3

Example Form for Simplified Catalog of Domains

Domains, as suggested in Exhibit 3, will be identified by a letter code, indicating dimension, and a serial number. Again there need be no system to the assignment of serial numbers. Likely codes for the various dimensions are:

- A: Age
- I: Industry
- L: Location
- E: Earnings
- $\phi$: Occupation
- T: Time

<table>
<thead>
<tr>
<th>Partition Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 10</td>
<td>No partitioning of A1</td>
</tr>
<tr>
<td>A 11</td>
<td>Partition of A1 in 5 year segments</td>
</tr>
<tr>
<td>L 20</td>
<td>No partitioning of L2</td>
</tr>
<tr>
<td>L 21</td>
<td>Partition of L2 by states</td>
</tr>
<tr>
<td>I 10</td>
<td>No partitioning of I1</td>
</tr>
</tbody>
</table>

(Continued on next page)
Partition Number | Description
---|---
11 | Partition of I1 by 2 digit SIC Categories
12 | Partition of I1 by 1 digit 1960 census categories
010 | No partitioning of Ø1
011 | Partition of Ø1 by 1 digit 1960 census categories
012 | Partition of Ø1 by 2 digit 1960 census categories

Exhibit 4

Example Form for **Simplified Catalog of Partitions**

The letter and first digit of the partition code (Exhibit 4) are the same as for the domain that includes the partition. The code "0" will be used to mean no partition of the domain. Catalogs of domains and partitions used with the full, detailed source survey will be similar, but stated with greater precision and detail.

<table>
<thead>
<tr>
<th>Source</th>
<th>Pages</th>
<th>Content Types</th>
<th>Domain Types</th>
<th>Partition Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17-34, 50,72-91, 112-114, 125,127, 129</td>
<td>Q1, Q2, Q4, S1, R5, q1,q2,q6</td>
<td>A1,6,5, L2,9, Ø1,2, I1</td>
<td>A11,12,61,50,52, L20,95, Ø10,21,22, I10,11,12</td>
</tr>
</tbody>
</table>

Exhibit 5

Example Form for **Preliminary Source Survey**
The Preliminary Source Survey (Exhibit 5) summarizes in non-detailed form the contents of the listed information sources. The purpose is to give a concise survey of the contents and the partitioning of the information in each source. These surveys will be used as a reference in making the full source survey, in preparing for computations, and in locating deficiencies in the information supply. Domain types and partition types need not both be given on the source survey sheet.

For the full survey of information sources, to follow and elaborate on the preliminary survey, the following forms will be used:
1. **Source List** (same as for the preliminary survey, see Exhibit 1)
2. **Catalog of Content Types** (same as for the preliminary survey, see Exhibit 2)
3. **Full Catalog of Domains** (like that for the preliminary survey, but expressed in more detail and with more precision, see Exhibit 3)
4. **Full Catalog of Partitions** (like that for the preliminary survey, but expressed in more detail and with more precision, see Exhibit 4)
5. **Full Source Survey** (see Exhibit 6)
6. **Survey Record** (see Exhibit 7)
The form for the Full Source Survey (Exhibit 6) is similar to that of Exhibit 5 for the Preliminary Source Survey, except in the following points:

1. Each occurrence of information is listed separately.

2. Full information about the domain and partition of each occurrence is given.

3. Units in which the information is expressed are specified.

<table>
<thead>
<tr>
<th>Source</th>
<th>Item Numbers</th>
<th>Pages Checked</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-10</td>
<td>1-50, 75, 81-93</td>
<td>All tables, pages 1-5, those marked on other pages</td>
</tr>
<tr>
<td>2</td>
<td>1-14</td>
<td>All</td>
<td>All tables included</td>
</tr>
<tr>
<td>3</td>
<td>All</td>
<td>All</td>
<td>This source contains no relevant information</td>
</tr>
</tbody>
</table>

Example of Survey Record

The Survey Record (Exhibit 7) serves to record the degree to which each source has been canvassed by the Full Source Survey. The
pages in the source which have been reviewed are indicated, along with the items that have been collected from these pages. A statement of the degree to which the source has been examined and to which its contents have been noted is also given. A similar form could apply later on to the coding and punching of information.

The most frequently occurring dimensions will probably be location, occupation, and industry. A sampling of some of the domains and partitions of these dimensions likely to be used is shown below.

<table>
<thead>
<tr>
<th>Location</th>
<th>Partitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domains</td>
<td></td>
</tr>
<tr>
<td>Full United States</td>
<td>By Region</td>
</tr>
<tr>
<td>New England</td>
<td>By State</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>By County</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupations</th>
<th>Partitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domains</td>
<td>Dictionary of Occupational Titles</td>
</tr>
<tr>
<td>All</td>
<td>U.S. Census 1950, 1 digit</td>
</tr>
<tr>
<td>All Civilian</td>
<td>U.S. Census 1950, 2 digits</td>
</tr>
<tr>
<td>Professional</td>
<td>U.S. Census 1960, 1 digit</td>
</tr>
<tr>
<td>Engineering</td>
<td>U.S. Census 1960, 2 digits</td>
</tr>
<tr>
<td>Skilled</td>
<td></td>
</tr>
</tbody>
</table>
Industries

Domains
- All
- All Manufacturing
- All non-Farm
- All Service

Partitions
- Standard Industrial Classification
- U.S. Census 1950, 1 digit
- U.S. Census 1960, 1 digit
- U.S. Census 1950, 2 digits
- U.S. Census 1960, 2 digits
- Dictionary of Occupational Titles

Many small variations in partitions will occur, and for proper processing must be made compatible or included in separate listings, whichever is appropriate.

Preparation of Formulas and Routines for Computation

The information collection procedures outlined earlier are the first step in combining the contents of individual sources of information to make a whole that is greater, in terms of the understanding it provides, than the sum of its parts. To this general end the following must be possible with regard to whatever information is collected:

a) to fill gaps in individual content categories through
   - interpolation
   - extrapolation
b) to condense and summarize information about individual content categories through
   - averaging
   - summing

c) to establish relationships among content categories in order to help fill gaps and to construct whatever new categories may prove useful

Additional computational capabilities that should also be included are:
   - Statistical procedures to suitably combine conflicting or overlapping information
   - Translation among partitions or domains
   - Discounting procedures
   - Normalization to satisfy constraints
   - Derivation of new partitions from sets of old partitions
   - Projections in terms of expected effects (e.g., technological change, urbanization) as a modification to purely extrapolative methods.

To illustrate that this prospectus includes within it the capabilities that have already been proposed, the methods of Reference 1, the first ISVD technical paper on forecasting, are shown schematically below in terms of the content categories discussed here.
Method 1

\[ Q_1(1966) \rightarrow Q_1(1980) \]

where \( Q_1 \) is employment specified by industry and occupation.

The above diagram indicates direct translation from 1966 to 1980 without use of further information. The equality relation in Reference 1 is a special case of this.

Method 2

\[ Q_1(1966) \rightarrow Q_2(1966) \rightarrow Q_2(1980) \]

\[ q_1(1966) \rightarrow Q_1(1980) \]

where the newly introduced quantities are defined as follows.

- \( Q_2 \) is total employment
- \( q_1 \) is distribution of employment among occupations and industries.

Method 2A

\[ Q_3(1966) \rightarrow Q_3(1980) \rightarrow Q_5(1980) \]

\[ Q_4(1966) \rightarrow Q_4(1980) \]

where the newly introduced quantities are defined as follows.
Q₃ is output by industry
Q₄ is output per worker
Q₅ is employment by industry.

Method 2B

\[\begin{align*}
Q₃(1966) & \quad \rightarrow \quad Q₅(1980) \\
Q₅(1980) & \quad \rightarrow \quad R₅(1966-80) \\
R₅(1966-80) & \quad \rightarrow \quad Q₆(1980) \\
Q₆(1980) & \quad \rightarrow \quad Q₆(1966)
\end{align*}\]

Where the newly introduced quantities are defined as follows.

R₅ is rate of growth 1966-80, nationally
Q₆ is regional or local employment by industry

Method 3A

\[\begin{align*}
q₁(1940) & \quad \rightarrow \quad q₁(1950) \quad \rightarrow \quad q₁(t) \\
q₁(1950) & \quad \rightarrow \quad q₁(t) \\
q₁(1960) & \quad \rightarrow \quad Q₅(t) \quad \rightarrow \quad Q₁(t)
\end{align*}\]
Method 3B

\[
\begin{align*}
Q_1(1966) &\rightarrow Q_2(1966) &\rightarrow Q_2(1980) \\
q_1(1966) &\rightarrow a_1Q_2 &\rightarrow Q_1(1980) \\
Q_3(1966) &\rightarrow Q_3(1980) \\
\end{align*}
\]

where like industries are grouped together, and

\[a_1\] is a coefficient relating employment mix and output within each group of similar industries.

Methods 4 and 5 are examples of adjustment or normalization of forecasts in terms of the results of other, related forecasts. They will not be discussed further here.

A comprehensive set of relations and formulas must be compiled and checked out, sometimes with alternate formulas for a given purpose. A list of these formulas will then be prepared, along with a body of rules for their use, including:

- terms and conditions of use
- form and type of input information needed
- form, nature, and possible use of the output, including an evaluation of its likely quality (e.g., accuracy, bias).

On this base a set of computer routines will be developed, to eventually constitute a specialized computer language to handle occupational and industrial forecasts.

To test and use this information system, not all available data
would be introduced at once. It would be coded in stages, as needed, and these would become part of the total supply of coded input data, to be combined and experimented with on an ongoing basis.

The following points, while not central to the above discussion, should also be kept in mind:

- We will want to allow for and include non-quantitative information, not for computation but for simple storage and recall.
- There should be a special information gathering and treatment program for short term information, using as sources
  - Job orders and similar local sources,
  - Newspaper advertisements,
  - Bureau of employment security materials,
  - Employment/unemployment figures.
- Computer output should be made to include statements of sources, of input quantities used, and of formulas used, to aid in checking the results.
- Computer output should indicate the information gaps found in trying to do computations, as an alert that these gaps will need to be filled.

There will be many lessons learned as the procedures suggested here are put to practical use. The present discussion is meant to be a point of departure, and should in no way limit the range or scope of future activity. Some important possible extensions of capability in forecasting are mentioned at the end of Reference 1. Of the many
possible directions that this work can legitimately take, the most easy to attain generally should be undertaken first. The more difficult ones will come after a foundation has been laid and the details of the procedures and analysis are better understood.
References

This list is meant to be suggestive only, and is in no way closed or restrictive.


