Data gathering for science and research

Techniky získávání dat pro vědu a výzkum

J. VANÍČEK

Czech University of Agriculture, Prague, Czech Republic

Abstract: Reasoning and argumentation in empirical sciences and research is based on row data and the intermediate and final structures, calculations etc., derived from the raw data. In this contribution, the short survey of the different techniques to gather raw data is given. The gain of the paper should consist in the manifestation of explicit limitation of usage this data for further utilization and deductions depending on the scaling type and validity problems during data gathering.

Key words: data gathering, categorisation, observation, measurement, scale type, data validation

Abstrakt: Usuzování a argumentace v empirických vědách a výzkumu se opírá o prvotní data, pomocné a výsledné struktury, výpočty atd., odvozené z těchto primárních dat. V příspěvku je uveden stručný přehled různých technik pro získávání prvotních dat. Přínos příspěvku má spočívat v objevení a uvedení explicitních omezení pro použití takto získaných dat pro další použití při odvozování závěrů, v závislosti na problémech validity dat a typu stupnic v kterých byla získána.

Klíčová slova: shromažďování dat, kategorizace, pozorování, měření, typ stupnice, validace dat

INTRODUCTION

The reasoning and argumentation in empirical science and research is generally based on raw data and structures, calculations etc. derived from the raw data. In the presented contribution we shall try to present different techniques to gather raw data. Anyway, all the interesting attributes of an empirical world cannot be described or measured in seconds and centimetres. Some phenomena of the real world shall be described using some more sophisticated formal structures than numbers, sometimes we shall describe the situation using many items, summarized into one figure. Not all operations with these formal items are invariantly related to the abandonment upon the concrete formal representation of attributes and therefore not all results of calculation are meaningful. Therefore, the measurement of them and scaling problems perform the essential role in the data utilization.

In the first part of this contribution, we shall try to present different techniques to gather raw data. In the second part, we shall indicate the main problems connected with measurement and scaling theory.

All the content of this contribution is focused on the conceptual-analytical researches, theory-testing researches, qualitative, interpretative and new theories creating researches and to artefact constructive researches according to the classification in Järvinen (1999). It does not cover the mathematical researches, mentioned in this monograph, because of the different character of the mathematical reasoning and the special position of mathematical and logical disciplines between the sciences.

DIFFERENT TECHNIQUES

The most typical data gathering techniques are observation, interview, questionnaire and written material. In addition, introspection video and voice recording, surveillance by a computer program, pilot tests or prototypes and artefacts can be mentioned. To use a combination of various data gathering techniques is generally recommended.

Observation

Everyone observes his environment. Anyone observes the actions of others. Scientists seek reliable and objective observations from which they can draw valid inferences. They treat the observation as a part of measurement procedure. Observations must be naturalistic, observers must be immersed in the ongoing realistic and natural situations and must observe behaviour as it occurs in the raw. Direct observation may be more reliable than what people say in many instances. It can be particularly useful to discover whether people do what they say they do, or behave in the way they claim to behave.

The contibution presented at the international conference Agrarian Perspectives XII (CUA Prague, September 18–19, 2003).

The following checklist (see Bell 1993, p. 119 – for the observation of actions of human behaviour only) is recommendable to follow:

1. Decide exactly what you need to know (list of all topics/aspects is recommendable).
2. Consider why you need this information (examine your list and remove any item that is not directly associated with the task).
3. Is observation the best way to obtain the information you need (consider alternatives)?
4. Decide which aspects you need to investigate (process, product, interactions, …?).
5. Request permission (official channels, individuals concerned).
6. Devise a suitable grid, checklist or chart (consult and adapt where necessary a published examples).
7. Consider what you will do with the data and whether the data contain the information concerned. (Is it likely to produce anything of interest? Will the data be the sufficiently complete to enable you to come to any conclusion?).
8. Pilot your technique and revise it if necessary.
9. Prepare carefully before observation (Draw the plan, indicate the arrangements and layout, prepare the etalons for measurement, if required).
10. Select a suitable place and period of observation.
11. Remember that no grid, however sophisticated, will tell the full story. (Try to place the event in this organizational context).
12. Analyse and interpret data, consider what the facts indicate or imply.

Interview

Interview can be described as a conversation between the interviewer and the respondent with the purpose of eliciting certain information from the respondent. The type of interview differs depending on the research approach used. In the theory testing approach completely formalised the interview, where interviewer behaves as much like a machine is recommendable. For the theory creating approach, a completely informal interview may be more suitable in order to get and present as all-round views as possible. The role of interview in the constructive approach differs depending on whether we have a building of new artefact or evaluation case. In the first case, the discussion between the developer and user is essential, in the second case, if the criteria to be used in evaluation are predetermined, the formalised interview can be appropriate, but if such criteria are still sought for, the informal interview is recommendable.

In the formalised or structured interview, the interviewer tries to stay as natural as possible and to interact with all respondents in the similar way. Based on research framework and hypotheses, the interviewer carefully plans and formulates questions. In the informal interview, the interviewer tries to gather the description of life-word of the interviewee with respect to the interpretation of the meaning of the described phenomena. Often such an interview is "semi-structured", it is neither a free discussion nor a strictly structured questionnaire. Establishing of an easy relationship by warming up before the interviewing session is useful in both types of interview.

The following main aspects to understand the qualitative research interview can be delimitated:

– Concentrating on the interviewee's life-word
– Seek to understand the meaning of phenomena in the interviewee's life-word
– Interview shall be
– Qualitative
– Descriptive
– Specific
– Presuppositionless
– Focused on certain themes
– Open for ambiguities
– Open to changes depending on interpersonal interactions.

Bell (1993, p. 99) presents a nice long checklist with 21 points for a good interview.

Questionnaire

Questionnaire in the paper or electronic format contains structured or unstructured questions, with the intention to be responded by the selected people. Questionnaire is mostly used for data gathering technique in the survey studies, one of the theory testing research approaches. Questionnaires with open questions can be also used in the theory/creating studies. The difference between structured and open questions corresponds to the similar difference between structured and informal interview and almost all instructions given for interview can be applicable also for questionnaires.

The practice for standardized questionnaires in survey research is based on the following assumptions:

1. The meaning of the questions shall be shared by the majority of respondents.
2. The respondents mostly understand the stimulus or phenomenon under investigation in roughly equivalent way.
3. The responses will be given in a manner, which allows the researcher to interpret and compare them.

Some recommendable criteria for question writing applicable also for interview follow:

a) Is the question related to the research problem or research objectives? All the items of the schedule should have some research problem relationship.
b) Is the type of question appropriate? If the information needed can be obtained also by observation, the more objective approach shall be selected.
c) Is the question clear and unambiguous? The ambiguous statement or item permits or invites differing responses resulting from alternative interpretations.
d) Is the question not a "leading question", which suggests answer?

e) Does the question demand knowledge and information that the respondent does not have? To control the invalidity of response due to a lack of information, it is wise to use information filter questions, i.e. "Whether a respondent knows what *xyz* is and means?"
f) Does the question demand a personal or delicate material of controversial nature that the respondent may resist? In the positive case, the special techniques exist to obtain such information, which is needed.
g) Is the question loaded with social desirability? With respect to sensitive behaviours, people will tend to under-report performance of social undesirability behaviours and over-report social desirable behaviours. When the behaviour of interest is not particularly sensitive, self-reports tend to be quite accurate.

The surveys that require users to evaluate or make judgments about information systems and their effect on specific work activities can produce misleading results if respondents do not interpret of answer questions in the ways intended by the researcher can be found n Hufnagel and Conca (1994).

Written materials and other documents

Such a sources, written, printed or recorded using various multimedia techniques, can be divided into primary and secondary sources. Primary sources are those sources, which came into existence in the period under research. Secondary ones are interpretation of events of this period, based on primary sources. Secondary sources may be originally collected for another purpose and in the process of accomplishing another activity.

To locate the possible documents can be sometimes a difficult task, which is often underrated in present time. The cyberspace including Internet now offers huge possibilities to easily collect text. But the Internet is rather a passive medium. The agent software can be utilised in data and knowledge gathering. This new area also causes new kinds of ethical issues for researcher. The amount of documentary material that you can study will inevitably be influenced by the amount of time that is available for the stage of research. It is not usually possible to analyse everything and so the researcher must decide what to select, based on the research problem.

The analysis of documents can be divided into external and internal criticism. External criticism aims to discover whether a document is both genuine and authentic. Internal criticism aims to analyse the context of the document.

MEASUREMENT

The measurement in the broad sense can be described as a method of describing the selected segment of a real word using some formal structure, which can be processed using mathematical operation and/or information technologies. Measurement is a powerful tool for the objectification and of any research. In the mostly frequented special case, the formal objects are interpreted as numbers (natural numbers, integers or real numbers) of the selected domain, so called measures. But often not all properties of numbers can be interpreted as properties, which are meaningfull in the selected segment of real world. This phenomenon is induced by the abandonment in the measurement. We can measure the length of the table by centimetres or by inches, the temperature in Celsius or Fahrenheit degrees and the achievement of the student during test using scale 1, 2, 3, 4 or using percentage from 100 to 0. Only such results derived by computation from measures (numbers) are meaningful in the real world, which are independently related to the abandonment of the measurement. The degree of the mentioned abandonment can be described using group of so-called admissible transformation of measures. This group identifies the scale type of measurement. The type scale used in making measurement on the variables will generally influence the choice of applicable techniques of data processing, in special statistical techniques.

Scale types

For the measurement in the narrow sense (using numbers), the following five types of scale are frequently used: *nominal*, *ordinal*, *interval*, *ratio* and *absolute*. These types are distinguishing basis of the group of admissible transformation of measure values, their algebraic properties and the types of mathematical operations permissible for each of them.

– A *nominal scale* assigns numbers only as an identifying label, to each object or set of objects under study. Assigning letters or different pictures instead of numbers serves just as well. Such measurement is in the fact only a categorisation. The set of all admissible transformations is the set of all one to one mappings in the domain of measures. The inequalities between the assigned measures (numbers) and the arithmetical operations between them have not any empirical meaning and it is necessary to avoid the attempt to interpret such operations in the real world.
– An *ordinal scale* assigns numbers to objects, which are rank-ordered with respect to some characteristic or our preferences. The classical example of ordinal scale type is the classification of pupils in the school or the structured questionnaire, where responses "strongly agree", "agree", "neutral" disagree" "strongly disagree" are coded by numbers 1, 2, 3, 4, 5, respectively (or qualifiers: extremely, quite, slightly, neither or equally, slightly, quite and extremely, combined with adjective sets are claimed to form such a scale. Generally it is not any argument for the presumption, that the differences in such steps are equal each to others. The group of all admissible transformations is the set of all strictly monotone functions. In the ordinal scale, the inequality symbol between the measures has an empirical meaning,

because it express the preference, but arithmetical operations have not empirical meanings.

– An *interval scale* is obtained if we can define unit of measurement such, that a difference of scale values has an empirical meaning regardless of whether it represents (say the difference between 10 and 9 is the same as between 7 and 6 and 6–5 etc.). The most well-known example of interval scale is temperature measurement using Celsius degrees, Fahrenheit degrees or Calvins. The set of all admissible transformation consists of the functions of the form $y = a \cdot x + b$, where $a > 0$ and b are arbitrary constants. In the interval scale, the arithmetical difference has the empirical meaning, but it is not meaningful for example to speak of one scale value being twice or three times as large as another.
– A *ratio scale* is in the fact the interval scale having further an absolute zero point. The only one freedom in the measurement in this type of scale is the selection of unit (for example meters, yards or versts). The set of all admissible transformation consists of the functions of the form $y = a \cdot x$, where $a > 0$ is arbitrary constant. The ratio scale is preferable to use if it is possible, because almost all arithmetical operations can be interpreted in the real world and have an empirical meaning.
– An absolute scale is a measurement with any freedom. Admissible transformation is only identity. The typical absolute scale type of measurement is the measurement by ratio or percentage of two values resulting by the same measurement in ratio scale.

It is very important to emphasize that the possibility to realize the measurement of the desired scale type depends of the properties of the real word segment, which we want to measure. The conditions, which must be satisfied for the possibility to measure in the desired scale, is expressed by the so-called representation theorems. For the ordinal scale type measurement, the respective representation theorem dues for Birkhoff and Milgram, which hedges the existence of such measurement on the relative week conditions in the real world (our preferences must be at least a week order and the measured set not "to large" – must contain a at most countable order dense subset). These conditions can be fulfilled in almost all practical situations except of such, when there exist entities or invents which are not comparable by our preferences.

For interval and ratio scale, the situation is not so straightforward. Primarily, we cannot speculate of such measurement if we have not any idea about the meaning of the difference in the real world (in the case of interval scale type measurement) or the concatenation operation between real objects (in the case ratio scale type measurement). But the existence of such concepts is not a sufficient condition. Many other not obvious conditions have to be satisfied. One, which is the key condition for the possibility of ratio scale measurement, is so-called Archimedean axiom which after a broad simplification says, that any given gap in our preferences can be overcame by sufficiently many tiny improvements. It is obvious, that the relevance of the Archimedean principles is not in place in many practical situations. See Vaníček (1999) and Vaníček (2000) for detail.

In the so-called "hard system research", for example the physical ones, measurement using ratio and interval scales are common. In the so-called "soft system research", for example in psychology, ecology and education, we often have to do with ordinal scale type measurement only, because it is impossible to realize the ratio or interval scale measurement. For example, it is still an open problem if the scores on well-constructed intelligence tests might be regarded as approximating an interval scale or not. Tatsuoka and Tiedeman (1965) estimate that the answer is yes. The author of this contribution point of view is that the answer is not. It is no doubt that ratio scale types are rarely, if ever, achieved in psychological, ecological and educational measurement.

Some quality aspects of measurement

The term *generalizability* refers to usefulness of a theoretical construct outside its limited domain of known observations. In one sense, strict *generalizability* is used in statistic based studies to indicate the probable mathematical relationships between observations of phenomena in a sample and the phenomena in the corresponding population. Baskerville (1996) used those definitions as a starting point when he tried to develop the concept of generalizability as a two-step process. But the second stage of his view is, in the mind of the author of this contribution, best interpreted as an application of the results from the first stage to the building an artefact. Baskerville does not seem to keep the building an artefact as a separate research problem.

Reliability regards to the extent to which observations by multiple researches studying the same phenomenon with similar purposes will yield approximately to the same results. Baskerville (1996) has distinguished three types of reliability. *Quixotic reliability* is the extent to which the particular technique of the observation yields an unvarying measurement. *Synchronic reliability* is the extent to which observations within the same time frame are similar. *Diachronic reliability* is the extent to which a particular technique of observation yields to the same measurement phenomenon exposed to a particular phenomenon at different points in time.

Validity regards the extent to which an observation measures what it purports to measure. Validity means that a theory, model or concept accurately describes reality. According to Baskerville (1996), validity can be analysed using different taxonomies. One such taxonomy deals with the semantic of the observational data. There are three types of validity with regards to the semantic of data: criterion validity, content validity and construct validity. Another taxonomy deals with the causal inferences and correlations of the elements of the theory. This second taxonomy is most commonly considered in the negative sense to find possible failures. *Internal validity* is the extent to which the causal analysis and explana-

tion offered by the theory reflect the reality at the moment of observation. *Content validity* is the extent to which the data from the object under study are related to the theory under test. *External validity* is the extent to which the causal analysis and explanations offered by the theory may be applied to similar phenomena. *Criterion validity* is the extent to which the data from the object under study will predict some important form of behaviour.

Multi-item measures

The measurement can concern either individual item or multi-item measures (vectors of measures). See Nunnallt (1978). The latter can be sometimes more recommendable. Churchill (1979) mentioned three reasons for this:

1. Indvidual item usually has considerable uniqueness or specifity in that each item tends to have only a low correlation with the attribute being measured and tends to relate to other attributes as well.
2. Single item tends to categorize objects (people) into relatively small number of groups.
3. Individual items typically have a considerable measurement error. Reliability tends to increase and measurement errors decrease as the number of items in a combination increases.

The author's recommendation is to add the further important reason.

4. If our intension is to map our preferences into the measures of the attribute of a given object and our preferences form only a partial order, not a weak order (no comparable objects exists), it is not possible to realize any ordinal type scale measurement by individual measure. But the multi-item ordinal type scale measurement is possible. The minimal number of necessary items is the topological dimension of the respective partial order (see Vaníček 2000).

Churchill (1979) suggested the *eight steps procedure* for developing multi/item measures. Steps 3 and 5 are the data collection steps, and hence we shall not describe them in more detail.

Step 1: Specify domain of the construct. After the researcher has performed a literature search and no acceptable measure was found, he/she should have a good reason for proposing an additional new measure. The researcher must be exacting in delineating what is included in the definition and what is excluded. For example Bailey and Pearson (1983) defined user information satisfaction as the sum of feelings or affective responses to distinguishable factors of the computer-based information product and services that are provided within organization.

Step 2: Generate sample of items. The researcher must generate items, which capture the domain as specified. Those methods that are typically productive in exploratory research are generally productive. Churchill (1979) also thinks recommendable to use literature search to find how combined variable has been defined previously and how many dimensions or components it has. At the beginning, the purpose is to develop a set of items, which tap each of the dimensions of the construct at issue. By incorporating slightly different nuances of meaning in statements in the item pool, the researcher provides a better foundation for the eventual measure. Near the end of the statement developing stage, the focus would shift to item editing.

Step 3: Collect data.

Step 4: Purify measures. The purpose of the purifying process is to analyse whether a certain item belongs to the domain of the construct or not. The split-half test, where items are randomly divided into two groups and correlation between the groups is calculated, also measures the goodness of the multi-item measure. There is one basic problem with using it: Different results may be obtained depending on how the items are split half. Crombach's alpha mean reliability coefficient, defined by

$$\alpha = (k/(k-1)) \cdot (1 - \sum_i \sigma_i^2 / \sigma_t^2)$$

where:

k is number of parts (items) in multi-item measure,
σ_i^2 is the variance of the item i, and
σ_t^2 is the total variance of the multi-item measure is recommendable to use for the selection of possible ways of splitting a set in half.

The factor analysis can be used to test whether the items are located to one factor or not. The items that much deviate from the rest of the set are excluded from the whole set of items.

Step 5: Collect data.

Step 6: Assess reliability of new data. The major source of error within a test or measure is the sampling of items. In the case sampling is appropriate, and if the items "look right", the measure is said to have face or content validity. Cronbach's alpha is the basic statistic method for determining the reliability of measure based on internal consistency.

Step 7: Assess construct validity. Churchill (1979) wrote that specifying the domain of the construct, generating items that exhaust the domain, and subsequently purifying the resulting measure should produce a measure, which is content or face valid and reliable. It may or may not produce a measure, which has construct validity. The construct validity is most directly related to the question of what the instrument is in fact measuring, what construct, trait or concept underlies a person's performance or score on a measure. To establish the construct validity of a measure, the researcher must determine the extent to which the measure correlates with other measures designed to measure the same thing and whether the measure behaves as expected.

Step 8: Develop norms. A raw score on a measuring instrument may be particularly not informative about the position of a given object on characteristic being measured because the units in which the measuring instrument is expressed may be unfamiliar. The position of the individual on characteristic can be estimated by comparing the person's score achieved by other people. The technical name for this process is "developing norms".

At the end of the contribution, let us to mentioned that multi-item measurement, that is measurement by vectors of numbers, not only by individual numbers, is not the last step in the possible useful abstraction. More sophisticated and complicated mathematical objects and structures can be used for the domain of possible measure values. Such structures can be also processed using the co-temporally information technologies. Such a generalized measurement and the theory of scale types in this general situation is still an open area for theoretical science and research.

CONCLUSION

Research approaches can be divided into a mathematical (and formal logical) approaches, oriented to study formal structures and approaches studying reality (for example natural science and social studies). This two approaches use different research method but there complete separation is not reasonable. Methods of mathematical analysis have to be applied to the data, which must be gathered from the reality. Data gathering cannot be realized without keeping appropriate rules. This contribution try to formulate main respective rules and regulations to meat right model of reality in input data for mathematical analysis and the main rules and principles for the interpretation of results obtained by mathematical analysis back into the reality.

REFERENCES

Bailey J.E., Pearson S.W. (1988): Development of a tool for measuring and analysing computer user satisfaction. Management Science, *29* (6): 519–529.

Baskerville R. (1996): Deferring generalizability: Four classes of generalization in social enquiry. Scandinavian Journal of Information Systems, *8* (2): 5–28.

Bell J. (1993): Doing your research project – A guide for first time researches in education and social science. Open University Press, Buckingham.

Churchill G.A. (1979): A paradigm for developing better measures of marketing constructs. Journal of Marketing Research, XVI: 64–73, Febr.

Hufnagel E.M., Conca C. (1994): User response data: The potential for errors and biases. Information Research, *5* (1): 48–73.

Järvinen P. (1999): On research methods. OPINPAJA OY. Tampere.

Nunnally J. (1978): Psychometric theory. McGraw-Hill, New York.

Tatsuoka M.M., Tiedeman D.V. (1965): Statistics as an aspect of scientific method in teaching and reaching. In: Gage (ed.): Handbook of research on teaching, Rand McNally, Chicago: 142–170.

Vaníček J. (1999): Why the indication of metric scale type is interesting for the software quality measurement? In: Prec. of the 4th IEEE International Software Quality Standards Symposium and Forum (ISESS'99) Curitiba, Add1: 13–27.

Vaníček J. (2000): Měření a hodnocení jakosti informačních systémů. ČZU Praha.

Arrived on 5th December 2003

Contact address:

Prof. RNDr. Jiří Vaníček, CSc., Česká zemědělská univerzita v Praze, Kamýcká 129, 165 21 Praha 6, Suchdol, Česká republika
e-mail: vanicek@pef.czu.cz