UNITED
NATIONS

E

Economic and Social Council

Distr.
General

ECE/CES/GE.41/2007/11
24 May 2007

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Group of Experts on Population and Housing Censuses

Tenth session
Astana, 4-6 June 2007
Item 3 (b) of the provisional agenda

CENSUS TECHNOLOGY: RECENT DEVELOPMENTS AND IMPLICATIONS ON CENSUS METHODOLOGY

The Use of High Speed Data Processing to Capture Census Data

Submitted by United States of America[*]

This meeting is organised jointly with Eurostat.

Summary

The Bureau of the Conference of European Statisticians (CES), at its meeting held in Washington, D.C. (United States) on 19-20 October 2006, approved the renewed terms of reference for the Steering Group on Population and Housing Censuses and the plan for future CES activities on population and housing censuses. The CES Bureau also agreed that the Steering Group would coordinate the work on the diverse types of meetings. The present paper was prepared on request by the Steering Group on Population and Housing Censuses, for presentation and discussion at the Joint UNECE/Eurostat Meeting on Population and Housing Censuses in Astana (Kazakhstan), 4-6 June 2007. The paper provides substantive basis for the discussion in the session of the meeting dedicated to "Census technology: recent developments and implications on census methodology".

[*] This paper has been prepared by United States Census Bureau at the invitation of the secretariat.

I. BACKGROUND

1. The United States Census Bureau (CB) is the primary source of basic statistics about the population and economy of the nation. The CB is best known for the Decennial Census, the Census of population and housing that it conducts every ten years. Between 6 March and 15 November 2000, the CB collected, captured, reviewed, processed, and tabulated data from over one billion pages of census forms. On 31 December 2000, the CB delivered state-level counts to the President. These counts are used to determine the number of seats in the House of Representatives allocated to each state. More detailed state-level counts are due by 31 March 2001 and are the basis for distributing over $180 billion to state and local governments.

2. The Year 2000 Decennial Census Data Capture System (DCS 2000) was the first attempt by Census to use commercially available electronic scanners and high-speed digital data processing to capture Census data. Some of the new elements of DCS 2000 were:

(a) Respondent-friendly forms;
(b) Request For Proposal (RFP) for Data Capture contract, incorporating innovative recognition test decks;
(c) Use of private industry (Lockheed Martin and TRW);
(d) Electro-optical scanners with built-in commercial paper feeders;
(e) Other commercial-off-the-shelf (COTS) components;
(f) Proactive printing Quality Assurance (QA) system;
(g) Optical Character Recognition (OCR) for write-in fields;
(h) Optical Mark Recognition (OMR) for check-box fields via direct scanning of forms;
(i) Primarily Key-From-Image (KFI) with limited Key-From-Paper (KFP);
(j) Large volume of (over 140 Million) forms processed;
(k) Full respondent names were data captured for the first time.

3. The DCS 2000 program Goals for data accuracy were to have at least 98 per cent accuracy for write-in fields captured by OCR, 96.5 per cent for KFI, and 99 per cent accuracy for check-box questions. In production, our scoring reported herein indicates that the OCR accuracy was 99.6 per cent, KFI was in excess of 97.8 per cent, and the check-box question accuracy was a little over 99.8 .per cent. The overall write-in field accuracy (for the merged data combining OCR and KFI) was 99.3 per cent. These important results are summarized in Table 1 below, with the standard errors included in parentheses.

Table 1: DCS 2000 Program Goals and Production Results

Quality Attribute	DCS 2000 Program Goals	DCS 2000 Program Production Results
Write-In Field OCR Accuracy	98.0%	99.6%(0.01%)
Write-In Field KFI Accuracy	96.5%	97.8%(0.05%)
Overall Write-In Field Accuracy	N/A	99.3%(0.01%)
Check-Box Question Accuracy	99.0%	99.8%(0.002%)

4. The results we obtained from *Production Data Quality Sampling* indicate that DCS 2000 production data quality significantly exceeded its Program goals. It is important to note that the Program Goals, (which were intended to include true data capture errors only), were achieved despite the fact that what we term “errors” in this report are really a combination of the following:

(a) True data capture errors;
(b) Human interpretation differences;
(c) Cases of unclear respondent intent;
(d) Residual “truth” errors.

A. Description of the Data Capture System for Census 2000

5. Automated data capture and the quality of the information produced lies at the heart of the DCS 2000 system. Many times in the image processing industry, products or systems claim automated character recognition rates of 99 per cent or higher. But these rates are frequently calculated on pre-processed character test decks that rarely give an indication of how a system will work in an operational environment. DCS 2000 can make the same accuracy claim, but at a question level and on live Census production data. Moreover, this rate is obtained with nearly 80 per cent of the data captured automatically. This level of automated capture did not come from simply a careful selection of commercial products or even by fine-tuning the individual OCR and OMR components of the system. These production statistics are the result of in depth tuning and complex integration of every component of the system.

6. For DCS 2000, form definition was the initial step of the tuning process. While DCS 2000, along with many forms processing systems in the industry, can define a form to run through the system in a matter of minutes, an in-depth analysis of Census rules, form characteristics, and respondent tendencies proved crucial to the production of a complete system definition that fully utilizes all optimizations that are designed into the form itself.

7. After DCS 2000 forms design and definition, the benefits of the tightly integrated components of the system became apparent. Normal variability of individual components was compensated for by strengths of subsequent commercial and custom developed components. Strengths of multiple components combined to produce results that none of the individual components could have produced on their own. And finally, these characteristics of the system optimized to produce more than just accurate data. They produced exceptional *Census* data.

8. As a very simple description of the system, the basic flow of the majority of forms is that they are first sorted in their envelopes at the sorting machines. Document preparation workers then manually remove the actual forms from the envelopes and assemble them into batches. For forms that contain multiple sheets (in booklets) the doc prep workers also remove the staples from the form to separate the sheets of the booklet. Once this step is completed, the forms enter the cluster and are scanned to produce a digital image of each side of each sheet of paper. This image then goes through a series of quality checks, mark recognition, character recognition, context checking, manual keying if necessary, and then the captured data is transmitted to the Census Bureau. As the last step of the process, the bar code on each form is wanded at the checkout stations in order to verify that the form has been properly processed through the system and that the Census Bureau has acknowledged that they have received the data.

B. Lessons Learned

9. There were some high-level lessons learned from Census 2000 which are now being implemented or studied for implementation in the 2010 Census:

a) Although the separation lines for the write-in fields worked very well for assisting respondents to print neatly, they had low visual contrast and were difficult for some people to see;
b) In the future, it will be important for Census to know data is correct while being processed, so new technology should be used to bring "production time" and "evaluation time" as close together as possible;
c) While the forms designs were a big improvement over 1990, there are many aspects such as multi-bank fields, multi-mark check-box question lay-out, etc that can be further improved;
d) The black boxes for check-box questions might be more pleasing to respondents if printed in a drop-out colour (with good visual contrast), and perhaps even improve check-box accuracy a slight amount.

II. PRODUCING AND ASSURING HIGHLY ACCURATE DATA CAPTURE

A Optical Character Recognition (OCR)

10. In the image processing industry, there are many OCR products that offer a wide range of capabilities and make claims of 99 per cent accuracy. However, in order to be chosen for the DCS 2000 system, the product needed to display the same architectural themes that are evident throughout the rest of the system. For this reason, an OCR product was chosen that not only produced high character recognition rates, but also provided internal contextual functionality such as trigram analysis and dictionary processing in order to make sure that production data would also be captured at equally high accuracy rates.

11. Over the course of not only this program, but previous ones as well, it has often been seen that an OCR product that advertised extremely high character recognition rates subsequently failed during evaluation tests that were performed. What is often the case is that the advertised rates were calculated on pre-processed test decks that were exclusively used for character recognition tests and did not require the use of a sophisticated segmentation algorithm. Because segmentation is one of the most important OCR processing steps and also one of the most difficult, excluding it from a test can produce significantly misleading results. Moreover, even with superior segmentation and character recognition algorithms, extremely high accuracy and acceptance rates are very difficult to achieve without further post processing.

12. So, in order to maximize the production of the system as a whole, the OCR product needed to follow the common design theme of properly performing its own task and allowing us as integrators to further cross check the system for errors in previous processing steps. For this reason, the fact that the product provides the following functionality made it an ideal choice as an OCR engine.

13. First, not only does it recognize characters with a high degree of accuracy, it also provides multiple choices for each character and corresponding bounding character coordinates.

This allows subsequent custom developed contextual processing to validate segmentation results as well as use an analysis of multiple recognition hypotheses in context and their probabilities of occurrence in order to further improve the results. Also, by providing a dictionary lookup capability as well as the description of the processing used to match or reject a word as a dictionary entry, the product allows even more opportunity for downstream analysis of the data during contextual analysis. Finally, because the product provides a vast array of definition parameters, it is also customized to treat each individual field with a high degree of detail and specificity, which will also maximize the accuracy and acceptance rates of the output.

B. Optical Mark Recognition (OMR)

14. On the surface, detecting marks in check-boxes may seem as simple as determining the number of pixels within a defined area and applying a density threshold. While this scheme will correctly determine the respondent's answer in excess of 90 per cent of the time, the DCS 2000 program required a significantly higher level of accuracy. In fact, because the desired accuracy was greater then 99 per cent, respondent mistakes often needed to be detected and handled as well.

15. As is the case throughout the DCS 2000 system, multiple levels of detection needed to be employed in order to achieve the desired results. In the case of OMR, the first level could be considered a simple pixel thresholding algorithm. At subsequent levels, more complex features were extracted from each mark in order to determine whether or not they are realistic enough to be considered the respondent's intended entry. Some of these features might include intersection points, relative density, and points of curvature. Each of these factors was evaluated in order to classify each mark as real or spurious. As an even further method of evaluation, multiple marks on a page were collectively evaluated in order to determine the respondent trend for providing a mark. This would provide additional information in cases were two equivalently "real" marks needed some sort of arbitration in order to determine the most probable representation of the respondent's intent. For instance, if the respondent answered all OMR questions with an "X" in the appropriate boxes, but on one question there was an "X" and a completely filled in box detected, the OMR algorithm could determine that the filled box was most likely a respondent mistake that was crossed out. In such a case, this spurious entry could then be removed.

16. However, during the evaluation of test data, one of the issues that arose was the political sensitivity of the race question in which multiple answers were to be expected. While these types of multiple answers were relatively rare, they were also difficult to arbitrate with extremely high accuracy. For this reason, in cases where multiple marks were detected (using very sensitive threshold settings) the marks were presented to two subsequent keyers for even more scrutiny. While this processing had little effect on the overall accuracy of OMR, the accuracy rates on this particularly sensitive area of interest were significantly increased.

C. Contextual Processing

17. While each of the previous OCR and OMR products achieve high levels of accuracy on their own, experience on previous projects as well as continuously developed experience on DCS 2000 showed that an even greater degree of accuracy and robustness can be obtained by using the contextual information that is inherent at not only the word and field levels, but also across different fields of the form.

18. The basic design of contextual processing is the quintessential example of the architectural theme in which a process validates and improves upon results of previous processing steps. The specific characteristics and tendencies of each of the previous steps from the time the respondent fills out the form, to the scanner, through registration, OCR, and then OMR, were taken into consideration in each contextual enhancement procedure.

19. The overall goal of this type of processing is to increase or decrease confidence in an OCR or OMR result based on contextual information. However, in certain instances, results may also be corrected. As the following descriptions of contextual analysis procedures will show, a large portion of contextual processing is spent on the detection and correction of errors due to noise (including noise generated by respondents) and segmentation errors. Experience throughout the DCS 2000 program as well as many previous image-processing programs have shown that these two problems are by far the most significant contributors to system errors. It should also be noted that the most effective validations have been shown to be the ones in which the variability in the field results are tightly constrained. This variability limitation can be applied to the character set allowed as output, the position of those characters, the quantity of those characters, and the relationship of those characters to either a defined result of even other characters found on the form.

1. Range validation Checks

20. Range validations are checks that a recognized number is within a specified range (Ex. month is 1-12). Often this range is tighter than the allowed values in order to take advantage of response probabilities and take into account expected frequency of certain error types. As an example, age field designs are 3 characters so the allowed value could be up to 999. However, the range of that field would be limited to somewhere around 99 since it would be rare to have an age over that limit and would avoid a third character (particularly a preceding '1') which could frequently be incorrectly added because of extraneous noise or a segmentation error. Because these two cases are two of the most probable sources of errors, this check is particularly valuable. Anything outside the specified limit would then be rejected and sent to a keyer for further review.

2. Multi-field numeric summation

21. This type of validation is essentially a test to make sure that two or more numeric results for a field add up to another or a specified value. One example of this is that a person's age plus their year of birth should effectively add up to the current year. This contextual edit is extremely valuable in that it helps to improve both error rate and reject rate for multiple fields. The reason that both of these inversely related rates are improved is that when the summation agrees, confidence levels can be increased – decreasing the reject rate. On the other hand, when the summation does not agree (usually because of one of the two most common sources of error – segmentation or character recognition) the confidences can be selectively decreased based on the most probable source of the discrepancy. This would then decrease the error rate. Contextual enhancements that simultaneously improve both of these rates are rare. However, because they are so effective, care should be taken when designing a form in order to maximize the opportunities for these types of multi-field, multi-improvement validations.

3. Character count validations

22. This validation is to make sure that a field contains a specified range of characters. It is most useful when the range is extremely limited. For instance, phone number area codes should always contain three digits. Any character quantity outside of this range is usually a sign of a segmentation error (one of the most frequent sources of error) and should be rejected.

4. Trigram analysis

23. This analysis is an examination of three consecutive characters of a word and their relative probability in a particular language. (Special treatment is given to the beginning and ends of words as it has been shown that these trigrams are even more limiting than internal word trigrams.) The results can be either a reduction in character confidences if the combination is rare or an OCR reassignment of character choices in order to improve the probability of the combination. For instance, if the last three characters of a word were recognized as “tbn”, a trigram analysis would show a low occurrence of this word-end in the English language and possibly reject the characters. However, if the second choice for the ‘b’ was an ‘o’, the results might be changed to become “ton”, a much more probable ending of a word.

5. List comparisons

24. Using an algorithm similar to a voting scheme, this type of contextual analysis is used to compare a list of similar fields in order to increase confidence of fields that agree. Also, if similar fields disagree and the voting scheme cannot determine a high probability result, the algorithm will consider rejecting uncertain fields. This scheme is particularly useful considering that many of last name fields and race type fields on individual census forms contain the same entries. As an example, if three of the four last names on a census form were “SMITH” and the forth was “SMITA”, it is highly likely that the last entry is “SMITH”. Depending on the statistics that are gathered as part of the algorithm, the last result may either be corrected or simply rejected.

6. Field associations

25. This type of analysis is used to detect extraneous entries in a set of related fields. For example, if a large percentage of a set of fields that describe a person is blank, but one field has some data in it, there is a high probability that the field actually contains noise. In this case, the data may be rejected or even deleted. Of particular note is that this algorithm takes both OCR and OMR results into account. In the case of OMR, the removal of noise errors due to such problems as scan lines created by dirty scanners is particularly difficult. This algorithm helps to detect and correct many of those types of errors, thus making the system as a whole more robust and less sensitive to problems generated at earlier processing steps.

7. Dictionary Results Processing

26. This processing will increase or reduce confidence of a word based on the status returned by the dictionary search performed during the OCR processing step. An extensive examination of the variations of respondent entries shows that nearly all dictionaries must be treated as open-ended. In other words, most dictionaries cannot contain all words. This is particularly true for names. While one might think that the dictionary could simply contain all spelling variations of

words, this has proven to be counterproductive in that the dictionary search algorithm has a more difficult task determining the corresponding dictionary entry for OCR results that are similar to multiple entries.

27. One of the most important lessons learned with dictionary results processing involves the relationship between the design of the respondent entry field on the form and the method that was used to match an OCR result to a dictionary entry. Because of the fact that the OCR fields are constrained, the probability of a segmentation error is reduced enough that additions or deletions of characters in order to match a dictionary entry are equally likely (if not more likely) to be the correct recognition of a spelling variation. For instance, if an OCR result from a constrained field was “PHILIP” the probability of the correct result being “PHILLIP” due to a segmentation error are similar enough to the probability that the original result was correct that a high confidence segmentation correction should not be the output result. While this situation may seem to only affect the minority of cases (i.e., the low frequency variants), the benefit is actually reaped during the creation of dictionaries. This benefit is that these low frequency, non-equal length variants of more probable words can be excluded from the dictionary search in order to increase the frequency of a high confidence match with the more probable variant. In other words, a dictionary search that finds “PHILLIP” (the more probable variant) can have a higher confidence in the match because it does not have to consider that “PHILIP” is a possibility. Moreover, the cases where the less-likely variant is actually the correct answer are not sacrificed as errors because the dictionary search correction of a segmentation error is rejected. The resulting effect is that the reject rate is significantly reduced while the error rate is maintained at a constant level.

8. Character Alignment with Constrained Field Boxes

28. One of the more difficult challenges that an integrator faces is one in which the characteristics of a product they are integrating into the overall system do not align with the goals of the system itself. While this following situation is very subtle, it is often the case with an OCR product and an automated data capture system. The reason is that most OCR engines are designed and built to have the capability to capture the widest variety of respondent inputs as possible. This means, that while they will often contain internal biases towards certain characteristics, they rarely constrain themselves to a high degree. When approached from the perspective of the OCR provider, this is done with good reason. In order to appeal to the largest customer base, they need to be as flexible as possible as well as extremely tolerant of less experienced users. This is often the case with recognition in a constrained field. While the OCR engine may bias its segmentation results based on the fact that single characters should appear in single constrained boxes, it might still attempt to output high confidence recognition results for respondents that have essentially ignored the constraint. Particularly since some of their customers will apply the same algorithms to unconstrained fields. The fact that the OCR engine is essentially forced into this situation leaves them vulnerable to a higher segmentation error rate on this subset of the universe of the response characteristics.

29. The alignment algorithm that was used in the DCS 2000 system was actually originally designed and implemented on a previous project. However, the original intent of the algorithm was to align OCR results from multiple OCR engines. Once that was done, those results were aligned with the constrained boxes. As was the case on the previous program, this algorithm has shown to be invaluable when it comes to detecting and correcting segmentation errors - one of

the largest sources of errors. The basic design of the algorithm is to use the coordinates of the recognized output characters of the OCR engine to align them with the coordinates of the defined constrained field boxes. Once that is done, the algorithm essentially determines if multiple characters appear within the same constrained box or if single characters span multiple boxes. If either of these cases is detected, a segmentation error or noise in the field is suspected and the results may either be rejected or corrected. The result is an extremely tolerant system that is less dependent on the segmentation design decisions of the OCR engine itself.

III. LARGE SCALE PRODUCTION LESSONS AND CHALLENGES

A. Forms Design

30. Of all the aspects of an automated data capture system; the absolutely most critical component of the system is the design and printing of the forms. A good form design can increase the stability, flexibility, error detection and recovery, and performance of the system. A poor form design can adversely affect all of these factors and subsequently increase the system cost exponentially. The experiences of DCS 2000 helped to emphasize these points.

1. Colour Choice

31. Final colour choices for the forms are extremely critical to the data capture process. This choice not only has to take into consideration the user-friendliness of the form, but also must take into account the wide variety of respondent writing implements as well. While it is preferable that each respondent uses a blue or black ballpoint pen, this cannot be guaranteed and in cases where enumeration is necessary, the use of pencils may be more cost effective. This was the case on DCS000 where it was much too costly to provide each enumerator with a ballpoint pen. For this reason, mechanical pencils were used. (Mechanical pencils were chosen over regular pencils in order to help avoid dull-pencil responses on the forms, which are more difficult to capture.) In situations like this, it is critical to consider form design colours that cleanly drop out and allow for aggressive contrast and threshold settings at the scanner in order to maximize the capture of such a wide variety of response characteristics.

2. Paper Specifications

32. Opacity and other paper specifications are also critical elements of the form design. These aspects of the design are often under-appreciated but are particularly important to the capture process. A particular example of this importance can be seen with forms that are two sided and where the design cannot be controlled to avoid printed type or entry fields on directly opposite sides of the paper. An additional example of the criticality of properly defined paper specifications is related to the accumulation of paper dust in the scanners as paper is fed through them. This problem can be minimized with proper paper specifications.

3. Document Integrity Handling

33. It would have been preferred to have a personalization id on all sheets of a form. However, that was too expensive so we needed to come up with a cost-effective way to print that still could reliably link all sheets together. In retrospect, if at all possible, avoid this added print cost and complexity by staying within practical limits of 16 page units.

4. Prototype Form Evaluation

34. It's never too early to evaluate prototype forms that have a reasonable chance to become the forms design baseline. Frequently, the form design process is heavily influenced by factors that can have adverse effects on the ability of the system to accurately capture data from the form (user-friendliness, etc.). In order to help assure that the automated data capture process is not compromised, the early incorporation of design goals that benefit data capture is a very effective strategy in order to help avoid conflicting objectives later in the process.

5. Forms Printing

35. Once the forms have been designed and the paper specifications properly defined, the production printing of the forms must take place. The aspect of this process that makes it particularly critical to monitor is that problems can quickly affect enormous quantities of forms. If left uncorrected for too long, the problems may become unrecoverable. Some of the lessons that were learned during this process are listed below:

(a) Bound or folded forms need to have critical dimensions defined from the centre or fold position so that a form at the minimum size and extreme fold position tolerance is not all reflected in one margin. In general terms, the most variable aspect of the printing process should be identified and all other specifications should consider that characteristic of the form;

(b) Constant on-site monitoring of the print vendor's quality assurance process by government representatives proved most essential to the quality of the Census 2000 product. Other Lockheed Martin forms printing experiences that did not have this constant monitoring have produced results that were not as consistent, further showing the need for a frequent on-site presence at the printer;

(c) Maintaining form design standards can help assure the consistency of the forms throughout the printing process. This can help to avoid very late or unexpected design changes.

*****