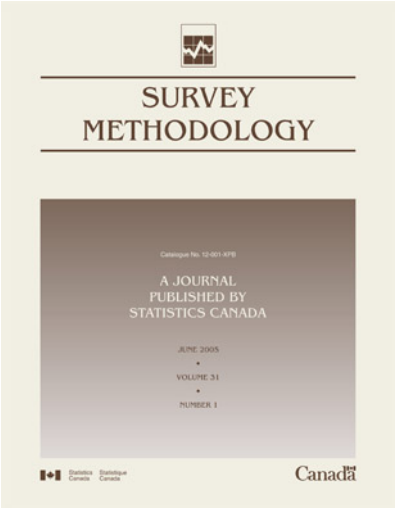
Catalogue no. 12-001-XIE

# Survey Methodology

December 2004




Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

| | |
|---|---|
| National inquiries line | 1 800 263-1136 |
| National telecommunications device for the hearing impaired | 1 800 363-7629 |
| Depository Services Program inquiries | 1 800 700-1033 |
| Fax line for Depository Services Program | 1 800 889-9734 |
| E-mail inquiries | infostats@statcan.ca |
| Website | www.statcan.ca |

## Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

# Survey Methodology

December 2004

Published by authority of the Minister responsible for Statistics Canada



April 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 30, No. 2, pp. 175-181
Statistics Canada, Catalogue No. 12-001

# Application of Quality Control in ICR Data Capture: 2001 Canadian Census of Agriculture

**Walter Mudryk and Hansheng Xie** [1]

Abstract

Intelligent Character Recognition (ICR) has been widely used as a new technology in data capture processing. It was used for the first time at Statistics Canada to process the 2001 Canadian Census of Agriculture. This involved many new challenges, both operational and methodological. This paper presents an overview of the methodological tools used to put in place an efficient ICR system. Since the potential for high levels of error existed at various stages of the operation, Quality Assurance (QA) and Quality Control (QC) methods and procedures were built into this operation to ensure a high degree of accuracy in the captured data. This paper describes these QA / QC methods along with their results and shows how quality improvements were achieved in the ICR Data Capture operation. This paper also identifies the positive impacts of these procedures on this operation.

Key Words: Data Capture; Intelligent Character Recognition (ICR); Quality control; Quality improvement; Statistical process control.

## 1. Introduction

The data capture of the 2001 Canadian Census of Agriculture was conducted between July and November 2001, using relatively new technology called Intelligent Character Recognition (ICR). This approach to data capture combines Automated Machine Capture which uses optical character, mark and image recognition, with Manual Capture by operators who 'key from image' using a heads-up data capture technique. The heads-up data capture technique is applied only to fields that can not be recognized by the optical system with a sufficiently high degree of confidence (that is pre-specified).

The ICR system offered many benefits to the data capture operation, in terms of resource savings and productivity gains. At the same time, accuracy became an extremely important consideration for processing a large number of documents since the potential for unacceptable levels of error existed at various stages of the process. In the literature, the quality of ICR applications has been studied by a few authors; see, *e.g.*, Kalpic (1994) and Pasley (2000), among others. Kalpic discussed the coding algorithm and the results for the 1991 Census Coding Operation in Croatia and Bosnia-Herzegovina, using intelligent optical readers. Pasley pointed out that the quality of a scanned image usually depends on the quality of the source document, the precision of the scanner, the skill of the scanner operator and the resolution at which the document was scanned. With quality improvement in mind, QA and QC procedures were built into the data capture operation for the 2001 Canadian Census of Agriculture to ensure a high degree of accuracy in this operation.

Quality Control activities for the ICR Data Capture Operation were focused in three main stages of processing, namely: document preparation, scanning calibration, and data capture of the questionnaires. This was done since each of these stages was dependent on one another and each had the potential to contribute significant errors down the line. Therefore, each component should ideally have its own control system.

It is the purpose of this paper to describe the QA/QC methodology and procedures associated with each of the main stages of the ICR Data Capture Operation, summarise the results obtained from their application and show how ongoing quality improvements were achieved in the ICR Data Capture operation.

## 2. Quality Program Overview

To better understand the rationale behind the QA/QC procedures, it is worthwhile to give an overview of their objectives and methodologies.

### 2.1 Objectives

The overall quality objective for this project was to measure, control and improve the quality of the entire ICR Data Capture Operation on a continuous basis. This would be achieved by implementing a series of QA/QC procedures at each critical stage of the operation. The specific objectives for each stage were as follows:

a) Document Preparation: to ensure that only highly readable documents would reach the scanning stage.

1. Walter Mudryk and Hansheng Xie, Business Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6.

b) Scanning Calibration: to ensure optimal machine set-up and calibration prior to the start of production.
c) Quick Capture (Machine Capture) and Quick Key (Manual Capture): to ensure a high level of quality of data capture during production.

### 2.2 QA / QC Methodologies

Each major stage of processing was operationally unique and therefore, had different quality requirements. As a result, QA procedures were applied to the Document Preparation operation, and QC procedures to the Scanning Calibration, Quick Capture and Quick Key operations. A flowchart is given in the Appendix, which shows the various stages of the ICR Data Capture Operation and exactly where these procedures were applied.

#### 2.2.1 Document Preparation

The document preparation operation was essentially divided into five sub-processes, specifically: sorting, transcription, batching, cutting and storage. This operation was responsible for preparing the questionnaires and associated batches for scanning by the ICR equipment and was performed manually by clerical staff. It included activities such as separating the contents of the received envelopes by document type (*Sorting*), re-transcribing damaged or illegible questionnaires (*Transcription*), grouping questionnaires into batches for registration (*Batching*), cutting the spine of each booklet questionnaire with an electric cutter (*Cutting*) and filing questionnaires in the archive (*Storage*). One of the most important aspects of this operation was the identification and isolation of problematic questionnaires so that they would not advance undetected to the scanning and data capture stages. These problematic questionnaires were labeled as '*outlier*' questionnaires since they had problems such as questionnaires being *X*'ed out or written over fields, extraneous markings, illegible entries, torn, crumpled or taped documents, *etc*.

The potential for error in this operation could lead to some problems being experienced at the scanning stage. It was felt that QA procedures would be appropriate to ensure quality at this stage since many of the clerical functions were also subject to various automated system cross-checks. The system cross checks ensured that the documents had a valid ID, correct number of pages, and that the pages, once cut, were aligned and in sequential order. The QA procedures consisted of a series of on-going random spot checks for each of the five sub-processes. The results of each spot check were recorded on a control form and summarized for the supervisor to identify if the work was being done correctly. Feedback would then be given to the individual clerk or group on a regular basis, and corrective actions would be taken when necessary. For example, if the work was not being performed well, some re-training would take place and/or an increase in the frequency of spot-checks was done until favorable results were obtained. If extensive problems were identified, the supervisor could also decide on the amount of re-work required, based on the seriousness of the problem observed.

For the sorting, batching, cutting and storage operations, the quality measure selected was '*percent of questionnaires in error*' (*i.e.*, in keeping with the assumptions required for a simple sampling unit). For the transcription operation, the probability of multiple independent errors occurring within a questionnaire was extremely high and therefore the quality measure selected was '*Defects per Hundred Units, DPHU*' (*i.e.*, in keeping with the assumptions required for a complex sampling unit).

#### 2.2.2 Scanning Calibration Check

Experience has shown that if the scanning equipment is not properly configured, the potential for generating poor quality images increases substantially. It is therefore imperative that the scanning equipment be optimally set prior to production and well maintained throughout the scanning operation. To ensure this, a QC procedure called the Scanning Calibration Check was developed to review the machine settings and calibration on an ongoing basis.

Since the equipment settings of the scanning system would tend not to fluctuate too greatly, it was felt that Statistical Process Control (SPC) methods would be appropriate for controlling this portion of the operation. This would essentially be an ongoing spot check of the calibration settings performed on a daily basis prior to the start of production. The calibration check consisted of re-scanning a test batch and comparing the results with the corresponding pre-benchmarked results for the same batch. The differences between the actual and expected results would be compared and error rates computed. These error rates were then plotted on SPC control charts to determine if the process was operating at an acceptable level. If this test batch failed, the scanning process would not be allowed to start production until the machine was re-calibrated and subsequently re-tested successfully.

In the Scanning operation, machine recognition could substitute wrong values when poor quality images are produced. Poor images could be the result of many factors such as dirty read heads, smeared optical windows, mis-alignment, mis-registration of fields, poor contrast / brightness levels, paper feed problems, *etc*. Since a specific quality standard was established for each field type, a separate $p$ control chart was used to evaluate the substitution error rate for each type (specifically, alpha, alphanumeric, numeric, tick boxes and bar codes). The acceptable quality standard for each field type was previously established on a

field type basis by the client area so therefore, the quality measure used was '*percent of fields in error*', *i.e.*, the substitution error rate by field type for each scanner.

Based on SPC control chart theory, a decision for each scanning calibration test was made as follows:

- If each of the sample error rates for the five field types was respectively lower than their corresponding upper control limit (UCL), it was concluded that the scanning system was functioning properly and was ready for scanning production.
- Otherwise, it was concluded that a problem existed with the scanning equipment, and corrective action must be taken before the start of regular production.

The test batches were constructed with minimum sample size requirements in mind for each field type, such that the producer's confidence level would be at least 95%. This was then used as a guide in selecting the actual questionnaires for each of the test batches. The *minimum size* was required for each field type in order to achieve the high efficiency of decisions in the scanning calibration test, while the *Producer's Confidence Level* referred to the likelihood that the scanning system would pass the test for that field type when the system was functioning at the acceptable target level. The Upper Control Limit for each field type was computed assuming a $+2\sigma$ variability. This limit is lower than the customary $+3\sigma$ Upper Control Limits since the scanning calibration check was designed to be more sensitive in detecting smaller shifts at start-up than during normal production.

### 2.2.3 Quick Capture and Quick Key

Once the questionnaires had been scanned, the system would produce a digital image of each field along with an interpretation of its value and an associated confidence level for its recognition. The actual data capture then consisted of two processes: Quick Capture and Quick Key. Quick Capture was the automatic recognition by the system of all field images whose confidence levels were above a pre-specified threshold value. Quick Key consisted of the heads-up manual capture (by keyers working on terminals) of field images whose confidence levels were below the pre-set threshold value.

Since under ideal circumstances, these two processes were expected to be relatively stable, the QC Procedures were again based on SPC principles and were developed to measure and monitor the quality of each of the processes. This QC approach consisted of a small sample check from the output of a sample of batches taken systematically over time and computing the error rates for each sample. These error rates would then be compared to rejection levels that were calculated by the system based on the expected quality standard and the sample size for that observation. A decision was then made as to the acceptability of each of these sample measurements relative to the expected quality standard for that process.

In the case of the Quick Capture operation, the machine may interpret a different value from the actual value for that field, and therefore, substitution rates were used to evaluate this process. These substitution errors are particularly serious since, if left unchecked, they may affect the recognition rate for many fields for a long period of time. In the case of the Quick Key operation, operators may make keying errors for many reasons such as lack of skill, poor training, fatigue, *etc*., and therefore, keying error rates were used to evaluate this manual process. For both of these processes, the quality measure was defined as '*percent of fields in error*', across all field types combined.

Within the two capture operations, there were two distinct categories for processing the scanned documents: *Regular* questionnaires and *Outlier* questionnaires. QC procedures were put in place for each category. A separate sample was required for each process, one for Quick Capture and one for Quick Key. The system could distinguish between Quick Capture and Quick Key fields in each sample questionnaire and maintain separate counts of these fields that had been captured under each process. These field counts eventually became the sample size for each sample. Each sample was then compared to its own *threshold rejection rate*, which was a function of the number of fields observed (*i.e.*, the effective *sample size*) and the expected quality standard or target for that process. A decision would then be made to accept or reject the sample. The threshold rejection rate was equivalent to the standard Upper Control Limit (*UCL*) that would be calculated on a standard *p* control chart. If the sample error rate exceeded this level, the process was rejected and the QC Reviewer proceeded to investigate and implement corrective actions as appropriate; otherwise the process was accepted.

The sampling was done on an individual scanner basis for Quick Capture and an individual operator basis for Quick Key. Some operators required more questionnaires to be sampled from time to time, and others less, based on their actual performance. Since the actual observations were based on samples, a customary $+3\sigma$ variability was permitted above the expected quality standard (*i.e.*, the centerline of a *p* control chart) for each process. The batch decisions for these sample observations were made by the system during QC verification and these results were then plotted on a *p* control chart for each scanner and operator, after the fact and updated weekly.

For a detailed description of these QA/QC procedures and their rationale, please refer to Mudryk, Bougie and Xie (2001).

## 3. Quality Improvements

Two essential elements were included in the quality improvement strategy for the ICR Data Capture Operation. These consisted of feedback of QA/QC results and the implementation of corrective and preventive actions when required. These two elements enabled various staff to play an active role in improving the quality of each process through the additional insight into the problems that were identified and through the subsequent corrective or preventive actions that were taken.

Using QC data analysis as the base, all processes were examined to determine if they were operating efficiently. QC meetings were held with operations staff on a weekly basis to review the ongoing progress of the entire operation. Problems that had impacted any of the processes were addressed and recommendations made to treat their root causes and prevent their re-occurrence. The involvement of operational staff in resolving these problems played an important part in facilitating quality improvements on a continuous basis. The following examples illustrate some of the more significant corrective actions that were taken during the operation that led to quality improvements at various stages.

**Example 1:** Filtering Process for Detecting Outlier Documents

During the first few weeks of production, it was noticed that some documents were causing a high concentration of errors from things like large X's across a page, 0's and dashes in various fields, *etc*. These documents were causing high error rates for both operations but especially for the Quick Capture process. Since these documents were very different from the majority of the regular documents, a procedure was introduced to sort these documents for special treatment and processing after the fact. Some documents in fact had to be re-transcribed at this stage prior to processing them by ICR.

**Example 2:** Adjusting System Settings for Scanning & Recognition

The highlights of the QC weekly summaries indicated that both scanners made errors frequently on Pages 3 and 14 of the questionnaires during the first few weeks of processing. An investigation was conducted and it was found that there was a template reading problem on Page 3 and the pre-set recognition threshold level for the numeric fields on Page 14 were set too low. After the system settings on both scanners were adjusted, the system showed substantial improvements in the scanning of these two pages.

**Example 3:** Retraining Operators with High Error Rates

During the keying operation, the QC results showed that certain keyers were experiencing above average difficulties with the 'key from image' process and that their error rates remained high for several weeks. Focusing on continuous improvement, these keyers were offered retraining on an ongoing basis. As a result, many keyers made significant improvements (week by week) in their keying performance.

## 4. QC Evaluation and Analysis

Throughout the operation, many QC reports, charts and estimates, were produced to provide information about the incoming and outgoing quality levels and to evaluate the output of each production process. These reports were used to analyse the quality of each process by week and across weeks.

### 4.1 Document Preparation

For each of the five sub-processes of the document preparation, individual QA procedures were applied at different frequencies and both corrective and preventive actions were taken on an on-going basis as dictated by the results. The information collected and the feedback that was provided as a result of these QA procedures helped significantly in improving the scanning, imaging, recognition and capture of the questionnaires. In the first few weeks of production, it was discovered from the QC results that problematic documents (*i.e.*, *outliers*) were causing most of the substitution errors (*i.e.*, machine errors) in the *Quick Capture* process. From that point on, a new procedure was introduced into the *Sorting* process of the Document Preparation operation to separate these documents for special treatment from the regular documents (*i.e.*, labeled them for subsequent 100% verification). In general, better quality documents reached the scanning stations while poorer documents were either re-transcribed or processed separately with the addition of post processes such as 100% verification.

### 4.2 Scanning Calibration Check

In an effort to ensure optimal scanner settings and calibration, a *Scanning Calibration Check* was initially conducted twice a day, and subsequently once a day, prior to production processing. Many test batches were scanned during the operation with a relatively high rejection rate encountered by each scanner. On average, approximately 2-3 tests per day (with corresponding re-calibrations) were required for optimising the set-up of each of the two scanners. This demonstrates the need for re-calibration between processing periods. It should be noted that some rejections occurred due to problems identified with the test batches which were fixed later on. This is definitely an area where some procedural improvement is required in the future.

Both scanners exhibited reasonably high variability during this test. The high number of tests required, high rate of rejection and high variability across processing periods for many of the field types demonstrate the need to calibrate the scanning equipment properly prior to production. Otherwise, the scanners could be inadvertently set up to produce poor images right from the start, which would make good quality capture very difficult. Once a test batch failed, problems were usually identified and subsequent maintenance and corrective actions taken. This included actions such as: re-configuring the scanning equipment, replacing old light bulbs, fixing software problems, cleaning dirty read heads, *etc*. Using this test, the scanners were able to be calibrated and maintained at optimum levels of performance, between production runs.

### 4.3 Quick Capture and Quick Key

For the ***Quick Capture*** process, over the entire 18 weeks of processing the ***Regular*** **questionnaires**, the overall weekly substitution error rates decreased steadily from 4.3% to 0.8%, resulting in a grand overall substitution error rate of 2.0% (across all field types) for both scanners. The substitution error rates measured during production were maintained very near the Target levels that were established for each field type. These were as follows: Alpha (2.1% relative to a target of 2.0%); Alphanumeric (3.2% *vs.* 3.5%); Bar Code (0.0% *vs.* 0.2%); Numeric (2.8% *vs.* 2.0%) and Tick Boxes (0.8% *vs.* 0.4%). In comparison, processing the ***outlier*** **questionnaires** had a much higher substitution error rate and greater weekly variability than the corresponding ***regular*** questionnaires (*i.e.*, ranged from a high of 22.4% to a low of 1.3%). Although the substitution error rate did tend to reduce substantially over time, it did remain relatively high throughout the process and was measured at 7.0% overall, which was significantly higher than the rate for ***regular*** questionnaires (*i.e.*, 2.0%).

For the *Quick Key* process, the keying error rate for processing the ***regular*** **questionnaires** was relatively high throughout the entire processing period (*i.e.*, mostly over 3%). This was partially due to the fact that this operation was a *heads-up* keying process and these keyers typically processed the most difficult cases. Over the entire 18 weeks however, the weekly keying error rates generally decreased from 5.6% to 1.6%, with an overall average of 3.4%. The keying was also subject to high levels of variability among operators, with individual error rates ranging 1.7% to 7.5%. It is interesting that keying the ***outlier*** **questionnaires** had a similar keying error rate to the corresponding ***regular*** process (*i.e.*, 3.4% *vs.* 3.7%) and ranged from a high of 5.7% to a low of 1.6%.

### 4.4 Estimates of Average Outgoing Quality

The primary purpose of the QA/QC procedures was to identify problems and to prevent them from occurring again. However, these procedures also had a corrective component in the sense that, errors that were discovered were always rectified. It is therefore possible to estimate the overall Average Outgoing Quality (AOQ) for the data capture component after the application of the QC procedures.

Estimates of AOQ were calculated for each of the two data capture processes. For a sampled outlier batch, all the questionnaires (*i.e.*, sampled and remainder) in that batch would be subjected to subsequent 100% verification, while for a regular batch, only the sampled questionnaires would be verified. This affects the calculation of AOQ since it can be assumed that the outgoing error rate for all verified questionnaires is 0.0%. The overall estimate for each component was based on the information obtained from both the regular and outlier documents, considering estimates of incoming quality and corrections made during verification. In the calculation, any documents reprocessed through either Quick Capture or Quick Key were included in the count.

Table 1 provides estimates of the AOQ for the *Quick Capture* and *Quick Key* processes.

**Table 1**
Estimates of AOQ for ICR Data Capture

| Process | No. Questionnaires in Population | No. Fields in Population | Estimated No. Fields Verified and Corrected | Incoming Error (%) | AOQ (%) |
|---|---|---|---|---|---|
| Quick Capture | | | | | |
| Regular | 273,818 | 21,248,277 | 170,249 | 2.01 | 1.99 |
| Outlier | 12,702 | 1,044,358 | 1,044,358 | 6.99 | 0.00 |
| Overall | 286,520 | 22,292,635 | 1,214,607 | 2.95 | **1.90** |
| Quick Key | | | | | |
| Regular | 281,502 | 6,376,020 | 234,253 | 3.41 | 3.28 |
| Outlier | 25,788 | 686,734 | 686,734 | 3.67 | 0.00 |
| Overall | 307,290 | 7,062,754 | 920,987 | 3.45 | **2.97** |
| Combined | | | | | |
| Regular | | 27,624,297 | 404,502 | 2.82 | 2.29 |
| Outlier | | 1,731,092 | 1,731,092 | 5.09 | 0.00 |
| Overall | | 29,355,389 | 2,135,594 | 3.24 | **2.16** |

It can be seen that the overall AOQ for the *Quick Capture* process was estimated at 1.90% and for the *Quick Key* process at 2.97%. This was down considerably from their corresponding estimates of incoming quality of 2.95% and 3.45% respectively. The overall AOQ for both processes was estimated at 2.16% (relative to an overall incoming error quality of 3.24%). It should be noted that the AOQ for *outlier* documents was assumed to be 0% since all *outlier* documents were subsequently 100% verified.

### 4.5 QC Summary

The above results clearly indicate the need for the QA/QC procedures at the different stages of processing. It also shows how they collectively contributed to controlling the outgoing quality and generating quality improvements into all phases of the ICR data capture operation.

The QC results clearly showed that the *outlier* documents had a greater negative impact on the *Quick Capture* process (*i.e.*, 7.0% substitution error rate) than the *Quick Key* process (*i.e.*, 3.7% keying error rate). This indicates that the filtering process for special treatment of *outlier* documents was an important step to take. The QC results also showed that if the documents were in good shape for scanning and the machines were well calibrated, the *automated* system was capable of capturing the data faster and with better quality than the manual *key from image* process. This is quite an important observation, since there are obvious savings implied with a corresponding improvement in data capture quality (*i.e.*, 2.0% *vs.* 3.4%). To the defence of the keyers, however, they did process the more difficult cases, thus partially explaining their higher error rates. Overall, it was estimated that about 77% of the fields were captured through the *Quick Capture* process and 23% were captured through the *Quick Key* process.

It should also be noted that the regular feedback of the QC information collected from the various stages of the ICR process was essential in identifying the root causes of many problems and in helping to resolve them. This provided the opportunity for many quality improvements to be generated into the various stages, on an on-going basis.

For a detailed description of these QA/QC results, please refer to Mudryk and Xie (2002).

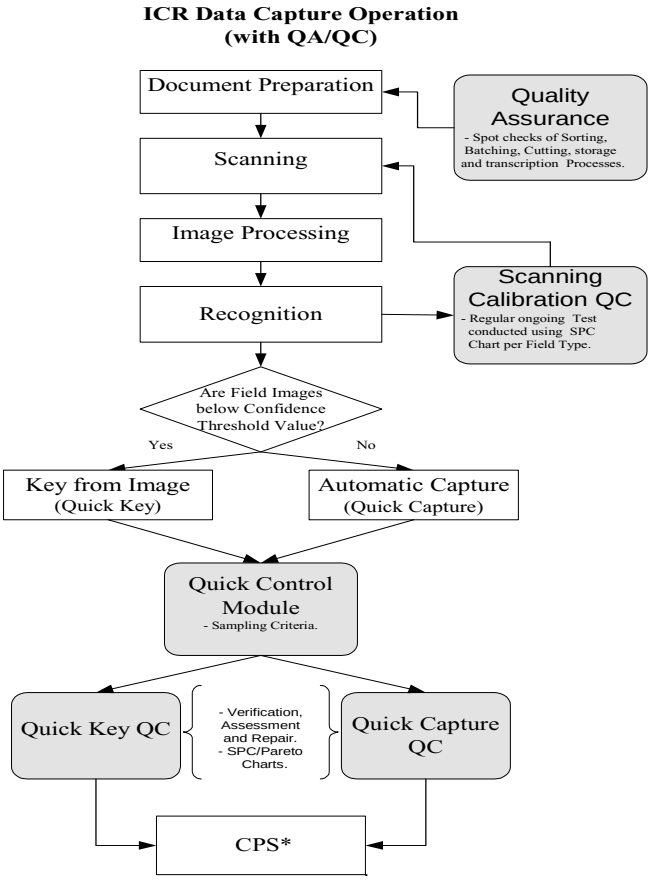
## 5. Conclusions

It is clear from the results obtained in this analysis, that the QA/QC procedures were extremely valuable and had a very positive impact on the entire operation. The QA procedures that were applied in the Document Preparation process were effective in preventing many poor documents from reaching the scanning stations and those that did were then labeled for special treatment and subsequent 100% verification.

The QC procedures were then able to optimize the machine set-up by applying the Scanning Calibration Check prior to production. Furthermore during production, QC samples were also able to identify problems with the automatic recognition and key from image processes, so that they could be improved as required.

In all cases, early warning signals were obtained from objective measurements at each stage of processing, and corrective and preventive actions were implemented as needed. Extensive feedback was provided to all stages of the ICR process on an ongoing basis from which continuous quality improvements were generated.

## Appendix




* CPS = Central Processing System.

## Acknowledgements

The authors are grateful to the Editor, to an Associate Editor and to an Assistant Editor for their detailed and constructive comments. They also thank Bob Bougie for many helpful comments.

## References

Kalpic, D. (1994). Miscellanea, Automated Coding of Census Data. *Journal of Official Statistics*, 10, 4, 449-463.

Mudryk, W., Bougie, B. and Xie, H. (2001). Quality Control of ICR Data Capture: 2001 Canadian Census of Agriculture. *International Conference on Quality in Official Statistics in Stockholm*, Sweden.

Mudryk, W., and Xie, H. (2002). Quality Control Application in ICR Data Capture for the 2001 Canadian Census of Agriculture. *Proceedings of the Joint Statistical Meetings,* American Statistical Association, 2424-2429.

Pasley, B. (2000). Web Exclusive: The Good and Bad of Scanned Images. Posted on the POB (Point of Beginning) website.