QUALITY ASSESSMENT AND CONTROL OF TOPOGRAPHIC DATA WITH APPLICATIONS TO HONG KONG 1:1,000 I-SERIES DIGITAL TOPO MAP

WONG KING FAI

M.Phil

The Hong Kong Polytechnic University

2016

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Quality Assessment and Control of Topographic Data

with Applications to Hong Kong 1:1,000 i-Series Digital Topo Map

Wong King-fai

A thesis submitted in partial fulfilment of the

requirements for the Degree of Master of Philosophy

February 2015

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material which has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____

(Signed)

Wong King-fai (Name of Student)

Abstract of thesis entitled "Quality Assessment and Control of Topographic Data with Applications to Hong Kong 1:1,000 i-Series Digital Topo Map" submitted by Wong King-fai for the degree of Master of Philosophy at The Hong Kong Polytechnic University in 2016

Topographic data depicts natural and artificial features on the earth surface, and is often merged using various data sources in order to reflect the changes of real world entities. The quality of topographic data may affect by many factors, such as uncertainties in source materials, data processing and data usages, etc. Data quality assessment and control are vital in managing topographic data which allows both data providers and users to obtain quality information of their data.

To assess the quality of topographic data, a clearly defined procedure is of major significant. This study analyses the principles of spatial data quality and proposes a procedure that includes four steps: a) create data quality objectives, b) establish data quality assessment plan, c) determine data quality result and d) report data quality result, for quality assessment. This procedure presents methods of obtaining internal and external quality requirements and introduces the principles of how to select appropriate methods for quality assessment.

In Hong Kong, topographic data is derived from several data sources and receives frequent updates to reflect rapid changes of city and supports town planning, construction and other purposes. To ensure the data fulfilling both internal and external requirements, a robust data quality controlling system is required. This includes the establishment of data policy, data documentation, data quality life control, error handling and data dissemination. This study reviews the topographic data

provided by Survey and Mapping Office (SMO) of the Hong Kong Special Administrative Region (HKSAR), and proposed a data quality controlling system that ensures the quality standards are maintained.

An experimental study, using the Hong Kong 1:1,000 i-series digital topographic map, is conducted to demonstrate the underlying principles. The significance of this study lies on three aspects: a) to analyze current status of quality assessment and control of topographic data, b) to develop a quality assessment procedure for digital topographic data and c) to enhance the quality controlling system of 1:1,000 Hong Kong i-Series Topo Map.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Wenzhong Shi for his contributions on my project. Dr. Shi has provided valuable suggestions and comments during my study. Without his support, I cannot complete my project and finish my research in the Hong Kong Polytechnic University.

Special thank is given to my wife, Anna and my sons, Matthew and Mike.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Topographic map is a picture of earth surface to help us understanding the spatial patterns, relationships and the environment that we live. The real-world phenomena are so complex; therefore the information showed on map is subjected to abstraction and variety of cartographic operations, such as classification, generalization and symbolization, to make it easier to understand. This creates differences between the contents of map and the reality, and arises the issue of spatial data quality.

Most digital topographic dataset are initially converted from paper maps using technologies like digitization and scanning, and then continuously updated by various data sources, such as remote sensing, ground survey and other data integration methods. Since the quality of some data sources may be undetermined, the quality of entire dataset is difficult to measure and describe. Although quality assessment is an expensive and time-consuming task, an accurate quality report is still worthy because it enables users to have more understandings about their products.

According to ISO 8402, the quality of a product or a service is defined as how much they fulfil the stated and implied requirements of users. This can be distinguished into internal quality and external quality. Internal quality refers to the level of conformance of a product to its requirements stated in product specification, while external quality refers to the level of concordance that exist between a product and user's expectation. Based on the principles stated in ISO 19113, both internal and external quality of spatial data can be described using quality elements and quality subelements. Examples of quality elements are: completeness, logical consistency, positional accuracy, thematic accuracy, and temporal accuracy.

To assess the quality of topographic data, clearly defined procedures are necessary to obtain consistent data quality result. Currently, national standards, such as National Standard for Spatial Data (FGDC-GPAS, 1998) and The International Organization for Standardization (ISO 19114), have developed methods and procedures for quality evaluation. These procedures mainly include five steps: 1) selection of quality elements, 2) sampling methods, 3) quality assessment methods, 4) statistical analysis and 5) quality reporting methods. Since some standards are established solely for evaluating positional accuracy of point features, the sampling methods and quality assessment methods are limited. This study proposes a new quality assessment procedure which presents the principles for selecting appropriate data quality assessment methods for point, line and polygon features.

Positional accuracy is one of important quality element that uses to describe the quality of spatial data. Point, line and polygon are major elements constituting vector-based spatial data. Positional accuracy assessment of point features has been well established and extensively adopted by contemporary positional accuracy standards such as ASPRS Accuracy Standards for Large-Scale maps (ASPRS, 1990), Geospatial Positional Accuracy Standard (FGDC-GPAS, 1998), and Australian Map and Spatial Data Horizontal Accuracy Standard (ICSM, 2009), etc. These standards recommended that the positional accuracy of spatial features is tested by comparing with the same features from an independent source of higher accuracy. This method assumes that the errors follow normal distribution and the sample sizes are sufficiently large. This study analyses the principles of positional accuracy assessment methods based on seven positional accuracy standards.

Line is another major geometry type of spatial data. Researchers (Shi, 1994, 1998, 2000; Goodchild M.F. and Hunter G.J., 1997; Heo J., Woo J., and Sang J.S., 2008) have proposed different methods to assess the positional accuracy of line features. These methods are reviewed in this study.

To maintain the quality of topographic data, an effective quality controlling system is essential to ensure the products fulfilling the requirements stated in product specifications and user's expectations. An experimental study, using Hong Kong 1:1,000 i-Series Digital Topo Map, is conducted to test the proposed quality controlling system.

1.1 Sources of Error in Topographic Data

Errors are unavoidable in measurements, so the true values are never known. In 1962, Greenwalt published a report introducing the principles of error theories for cartographic applications. According to this report, errors are classified into three categories named: 1) blunders, 2) systematic and 3) random. Blunder, is also known as gross error, which is mistake caused by careless observation. They are usually large and easily to deal with by repeated measurements. Systematic errors are a constant value that generally attributable to known circumstances. They conform to mathematical law and can be calculated and applied as a correction to the measured quantity. Careful calibration of equipment is an essential method to control systematic error. Random errors are those remaining after blunder and systematic errors have been removed. They are predictable and assume following normal distribution and obey the law of probability.

In digital geodatabase, the possible sources of errors may be derived from data capture, data processing and data usage.

1.1.1 Data Capture

Topographic data solely acquires from single source. It is usually integrated with various data sources with diverse data capture methods. These methods can be classified into primary and secondary data sources. The classifications of geographic data capture methods are shown in Table 1.1.

Table 1.1 Classification of geographic data capture methods

Data Capture Methods	**Raster Data**	**Vector Data**
Primary data capture methods	a) Digital satellite images b) Digital aerial photographs	a) GPS measurements b) Field survey c) Laser scanning d) LiDAR survey
Secondary data capture methods	a) Scanned maps or photographs b) Digital Elevation Models (DEM) from topographic map contours	a) Direct digitization from paper maps b) Direct digitization from photographs or images c) Interpolated data from models

1.1.1.1 Vector Digitization

Point is a basic unit in vector data model. Line is composed of sequenced of points, while boundary of an area is delineated by lines. To digitize a point feature, it is necessary to record its location in x, y coordinates. Attributes about the feature, such as feature code, is entered by operators. Some point features, such as spot heights, do contain z coordinate, the elevation information is inputted manually. To convert a graphic line into digital form, no matter straight or curved, it is necessary to break it into series of points. Two digitization methods can be used: a) on-screen digitization for scanned raster map and b) table digitization for paper map.

Digitization can be done in either point or stream mode. For point mode, a single point is passed to the host computer using digitizing cursor. It is useful for digitizing point features and points on line features, such as building outlines and roads. Using this to digitize curved features is inconvenient, because points on line must be digitized one by one. For stream mode, the digitizer hardware invokes an automatic sampling process and generates points automatically. Sampling can be according to distance or time, but time sampling is much more common for map digitization. When using stream mode, the operator needs to press the digitizing button only at start and end for a line. The points in between the line will be generated automatically when the operator tracks along the line.

The accuracy of vector digitization is depended on the quality of the original manuscript, skill of operator, and the precision of devices. The positional accuracy of original manuscript is described by national standards. For example, published map meets the position accuracy standard depicts 90% of well-defined points within 0.5 mm of their correct positions at map scale (NMAS, 1947). The original manuscript might also be distorted. Paper expansion and contraction due to humidity can span a range of ±0.05%. When a feature shows on paper map has 0.5 m length, it can amount to 0.25 mm. These distortions can be corrected through the registration process which calculates the transformation between the digitizing table and the grid coordinate system. There are three transformations commonly used in table digitizing: similarity, affine, and projective. Similarity preserves shape, affine preserves parallel lines, and projective preserves straight lines.

1.1.1.2 Ground Survey

Ground survey records general topographic information, such as buildings, roads, rivers and elevations on ground, in an appropriate scale. Detail survey is conducted based on control points which the positional information (x, y and z coordinates) are known. From a control point, horizontal angles, distances, and vertical angles to the details are measured, and then the coordinates of these details can be calculated. Total station is common instrument that used in ground survey because both distance and angles can be measured and stored in the instrument.

At present, a high-performance total station has an accuracy of 0.5" and at least 1 mm ± 1 ppm for angular and distance measurements respectively.

1.1.1.3 Satellite Positioning Technology

Satellite positioning system encompasses three major components named: ground control segment, space segment and user segment. The ground control segment comprised a master/control station, transmission stations, and monitor stations. The positions of the monitor stations are accurately known. They continuously monitor and receive signals from all satellites in view. The data collected at the monitor stations, including the predicated satellite orbit, the health status of the satellites and their correction parameters, are transmitted to the master/control station to form the navigation message. The navigation message is then sent to the transmission stations to be uploaded to all satellites, and broadcast from the satellites to the users.

The space segment consists of satellites that continuously broadcast measurement signals and navigation messages to users. They continuously transmit two L-band signals propagated in sinusoidal manner, referred to as L1 and L2 carrier phase signal. The L1 and L2 carrier signals also act as mediums to transmit the series of "0" and "1" code sequences to users by signals modulation techniques.

The user segment includes a satellite positioning receiver which contains both hardware and software for receiving, decoding, storing and processing collected data to determine the receiver's position. These data include the navigation message and distance measurements.

The positional techniques for surveying are based on differential GPS (DGPS) which is either code or carrier phase measurements on at least two stations. They are then classifies into several modes: static, rapid or fast static, stop-and-go, reoccupation, and kinematic. These modes differ mainly in their time span of observation, the time required for ambiguity resolution, and whether the receiver is static, moving, or a combination of the two throughout the observation period. Currently, mobile GPS receiver systems are widely used in positioning. They provide real-time location information for users.

The accuracy of satellite positioning technologies is closed related to the positional modes and measurement types that are being used. The accuracy of standalone GPS using pseudo range measurement is around 10 to 20 meters. A DGPS uses a reference station to reduce errors in measurements can reach meter level. The GPS Real-Time Kinematic (RTK) can be used in real time with positional accuracy of centimeter level, and the carrier phase measurement can reach millimeter level.

1.1.1.4 Airborne Laser Scanning

Airborne Laser Scanning system is widely used for the acquisition of dense and accurate terrain model on large area. A typical Light Detect and Ranging (LiDAR) system includes three main components named: GNSS system, INS unit and LASER unit. The GNSS system and INS unit provide positional and attribute information of the aircraft, while LASER unit provides range information from sensor's firing point to ground points. In addition, the reflective intensity (*l*) of the target is also detected and recorded, so the result of laser scanner usage is four-dimensional coordinates (*x, y, z, and l*) for each target point.

According to E.P. Baltsavias (1999), the accuracy of 3D coordinates is mainly affected by: a) range accuracy, b) positional accuracy, c) attitude accuracy, and d) time offsets. The range accuracy refers to the ability to measure the time interval between the same relative position on transmitted and received pulse. This is limited by noise, signal strength and sensitivity of detector, and shortness and reproducibility of the transmitter pulse. Positional accuracy refers to the quality of DGPS post-processing. This includes the GPS receivers, satellite constellation during the flight, distribution of ground reference stations from aircraft, misalignment between GPS and INS, and INS and laser scanner, and the accuracy of the laser beam direction. Attitude accuracy is the quality of the INS which will increase with the flying height and the scan angle. For accurate LiDAR survey, orientation, position, and range are required to be taken at the same time. If there is time offset, or this is not known precisely, it will cause a variable error.

The result of LiDAR survey is usually in WGS84. The transformation from WGS84 to the local coordinate system, including corrections of geoid model, is significantly affected the accuracy of LiDAR survey. The point accuracy of Airborne Laser Scanning can reach decimeter level, depending on above factors, data processing and ground conditions.

1.1.1.5 Image Acquisition

Images can be classified into two types: a) analogue images and b) digital images. In most GIS applications, it is necessary to have imagery in digital form, so analogue images need to be scanned before they used.

Aerial photography and satellite image are widely used in various GIS applications. For aerial photogrammetry, the camera is mounted on a plane. The ground surface is photographed by flying in a regular pattern on overlapping strips. This is called a block of photographs. The amount of overlap along the strip is called endlap which is usually expressed in terms of a percentage. For stereoscopic viewing, the aerial photographic shall at least have 60 % endlap. The distance between adjacent strips that form a block is called the strip width. The difference between the strip width and the ground coverage is called sidelap. To ensure continuity between strips, 30 % sidelap is preferred. Photogrammetry is the technique to extract reliable position information from overlapping images and blocks of images, and then form the geo-referenced images. This includes several processes, such as aerial triangulation, interior orientation, exterior orientation, stereo mapping and the production of DEM and digital orthophoto.

Space remote sensing acquires information about the earth's surface on a systematic, repetitive, global and multispectral basis. With the contemporary technologies in satellite image sensors, images with medium to high resolution covering a wider range of the electromagnetic spectrum can be acquired periodically. Atmosphere conditions, ability of sensor, and resolution of images are major factors affecting the accuracy of remote sensing's products. The positional accuracy of remote sensing images is continuously improved: from 60 m for Landset, 30 m for TM, 10 m for SPOT, 5 m for SPOT-5, 1 m for IKOMOS, 0.67 m for QuickBird, and up to 0.05 m for WorldView-1 images. The spectrum resolution has been developed from the Multi-spectrum of several bands to the Hyper-spectrum of over 100 bands.

1.1.2 Data Processing

Data processing errors mainly occur in data conversion, data integration, generalization, data editing, coordinate adjustment and other geo-spatial operations that are available in GIS software. These errors influence the positional accuracy, logical consistency, attribute accuracy and thematic accuracy of spatial database. Hunter G.J. (1991) presented about 70 data processing functions which can be classified into nine categories: a) coordinates adjustment, b) feature editing, c) attribute editing, d) Boolean operations, e) display and analysis, f) generalization, g) raster/vector conversions, h) data input and management, and i) surface modeling. These processing functions cause changes in both geometry and attribute of spatial features that are difficult to predict and assess if no rules have been defined before execution.

Mismatch of diverse spatial data sources is a typical error in spatial database. Edge matching or rubber-sheeting is a data processing function to remove such errors. In this process, it is necessary to prescribe the maximum allowable distance for edge matching, such that the influence on positional accuracy is minimized. Modern GIS software provides data processing functions to remove geometry errors, such as undershoot, overshoot, pseudo-node, broken line and artefacts line, etc. These functions may affect the accuracy of digital database if controlling processes are not undertaken.

1.1.3 Data Usage

Spatial data shared on internet rarely contains complete and meaningful quality information and thus leads to misuse. For example, OpenStreetMap are collaborative web-mapping agency that collects spatial data and delivers online map service that is freely available in internet. OpenStreetMap obtains data from government agencies and encourages contributors update the map using their local knowledge. Users are allowed to download map data via internet, but there is no quality information. Therefore, users are difficult to decide whether the data have sufficient quality to support their intended applications.

1.2 Principles of Spatial Data Quality

Starting from the early of 1940s, national mapping agencies (NMAS, 1947) and researchers (C. Vector Wu, 1994; Guptill S.C. and Morrison J.L. 1995, Goodchild M.F. and Hunter G.J., 1997) had raised the issues of spatial data quality and began to establish standards and methodologies to deal with problems of quality assessment and reporting.

In 1992, spatial data quality elements were first introduced in Spatial Data Transfer Standard which included lineage, positional accuracy, attribute accuracy, and logical consistency. They are then incorporated in the Metadata Standard under National Spatial Data Infrastructure (U.S.). In 1995, Guptill S.C. and Morrison J.L., from the International Cartographic Association (ICA) working group, published a book entitled “The Elements of Spatial Data Quality”. They added two quality elements named temporal and sematic accuracy. The temporal accuracy refers to the accuracy related to temporal aspects, and semantic accuracy refers to the quality of the definition of the entities and attributes in the spatial database.

In 1998, technical committee 287 of the Comité Européen de Normalization (CEN/TC287) developed the European pre-standard “ENV12656” which listed several quality elements that used for transferring data quality information between users. CEN/TC287 working group added two spatial data quality elements into the list: usage and homogeneity. Usage describes the intended use of dataset which allows potential users to understand the previous applications of dataset. Homogeneity indicates how much a quality elements are valid for the whole dataset or valid for only parts of the dataset. At the same time, ISO started standardization and CEN/TC287 dissolved into ISO/TC211 and ENV12656 ended. ISO/TC211 integrated the experiences of European and North American and other countries like Asia, Australia and South Africa. The objectives of ISO/TC211 are to establish a set of standards for handling of spatial data and other relevant services such as management, acquisition, processing, analysis, access, presentation and transfer, etc.

Until 2001, International Standardization Organization (ISO) working group started publishing several standards (ISO 19113, IO 19114 and ISO 19115, etc.) which defined spatial data quality in terms of data quality elements and data quality overview elements. Data quality elements include completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy while data quality overview elements include purpose, usage and lineage. These standards aim to establish procedures for evaluating quality of geographic data and define methods for presenting the quality of geographic information and products. ISO 19113, ISO 19114 and ISO 19115 are standards that commonly used to assess and report spatial data quality.

1.2.1 Positional Accuracy

Spatial data, no matter it is point, line, polygon or pixel, is the representation of position of ground features which includes horizontal values and vertical values in a specific coordinate system. According to ISO 19113, positional accuracy can be classified into absolute (external) accuracy, relative (internal) accuracy and gridded data positional accuracy and their definitions are stated explicitly in this standard.

Spatial data often captures by various methods and under different degree of generalization, so the positional errors may derive from: a) measurement error which included random error, systematic error and gross errors; and b) generalization error which included the uncertainty in the boundary recognition from the ground features and the ability of abstraction. Measurement error can be classified as error in the spatial data capture process which contained instability of the observer, statistical deviation, limitations of measurement instruments and unfavorable observation conditions, etc. Generalization error, as mentioned by Aalders H.J.G.L. (2002), is the amounts of generalization in the terrain during measurement by defining the feature's boundary to be surveyed.

Positional accuracy is determined by statistical methods and specified by Root Mean Square Error (RMSE), value of error (μ) or standard deviation(σ). Various methods for assessing positional accuracy of geographic data have been proposed by researchers such as Goodchild and Hunter (1997), Shi W.Z., Liu W. (1994), Heo, Woo and Sang (2008), and some of them are reviewed in section two.

1.2.2 Attribute Accuracy (Thematic Accuracy)

Geographic data may contain different types of attributes such as building name, address, area, population, type of soil, type of facilities, etc. Their accuracy can be measured in four kinds of scales: nominal, ordinal, interval and ratio. The nominal scale is a qualitative attribute, for example, building name. This type of data often expressed in textual format, and is assessed through comparison between the dataset and reference data. The ordinal scale is also a qualitative attribute, for example, classification of road. Depending on product specifications, some features were classified into different orders based on their size, length, rank or nature, etc. To assess the quality of ordinal attribute, clear statements stated in product specification can be used as conformance level for quality assessment. Both interval and ratio scales are quantitative attribute, for example, temperature, population and distance between two places. This type of data is usually described by symbolizations on map and is assessed through comparison between the dataset and reference data.

1.2.3 Temporal Accuracy

Real world entities are changed from time to time. Spatial or non-spatial information stored in dataset is only a representation of entities at a specific time. As more and more users considered that historical map, such as aerial photograph or remote sensing image, is useful information to record the history of a country or some special events, temporal accuracy must be addressed in GIS industry.

Guptill S.C. and Morrison J.L. (1995) defined temporal information of spatial data as: a) event time; b) observation time; and c) transaction time. Event time refers to an event occurs, changes or disappears. Observation time refers to an event being recorded while transaction time refers to an event added to database.

1.2.4 Logical Consistency

Logical accuracy refers to how well the data itself is consistent with each other. According to ISO 19113, logical accuracy is divided into domain consistency, topological consistency, conceptual consistency and format consistency. Geographical data may contain inconsistent which derived from data updating, data conversion, data processing or miscoding. Silver polygon, invalid geometry, edge mismatching, duplicate features and miscoding are typical examples of data inconsistent. Nowadays, most GIS software is capable of detecting and fixing these errors.

1.2.5 Completeness

Completeness is classified into commission error and omission error, where they are measures of excess and absence of data in a dataset. Usually, the data capture requirements of real world objects, such as minimum size, width or radius, are stated in product specifications; completeness is described how well a dataset fulfill these requirements. According to ISO 19114, quality of completeness can be reported by number or percentage of commission and omission items.

1.2.6 Lineage

Lineage describes the history of a dataset which includes data sources, data conversion methods, data acquisition methods, and processes that used to create and update the dataset. Lineage contains two major components:

a) Source information which describes the data sources of a dataset; and

b) Process steps which describes the processes used to update the dataset whether continuous, periodic or the lead time.

1.3 Objectives of this Study

Data is nearly the heart of all GIS functions. During the past thirty years, the issues of spatial data quality have become a significant area of focus for many national mapping agencies and researchers. They have developed theories and methodologies in modelling, assessing and reporting quality of spatial data. This includes approaches in evaluating positional accuracy of points, line and polygon, and understanding of how errors are propagated through data capture, data processing and data usage. International standards and theories have been published to deal with problems in sharing spatial data that along with comprehensive metadata and meaningful quality information. The ultimate goal is to help data providers and users to assess the "fitness for use" of their dataset. This study analyzes various data quality evaluation methods and reviews seven international standards based on the principles stated in ISO19113.

Vector topographic data is one of major components in contemporary GIS applications and is being widely shared, transferred and used by GIS specialists and general users. A simple and efficiency procedure to evaluate and report quality of topographic data allows both data providers and users to decide whether the dataset has fulfilled the requirements of product specifications and intended uses. This procedure must encompass principles of how to select suitable sampling methods, quality evaluation methods, and quality reporting methods. This study proposes a procedure for assessing the quality of topographic data.

Data may go through a number of processes, such as data capture, editing, storage, conversion and manipulation, etc., before being supplied to the customers. Within these processes, data quality must be checked by quality control and quality assurance measures such that the final products are satisfied the product specifications and fulfilled the requirements of potential applications. To maintain the quality of topographic data, quality control and assurance measures shall apply at all stages of data production process to ensure data integrity is achieved. This study adopts the Hong Kong 1:1,000 i-Series Topo Map as an example to demonstrate a quality control system for production of vector topographic data.

The objectives of this study are:

a) To analyze the principles and methods that used for data quality assessment and control of topographic data.
b) To develop a quality assessment procedure for topographic data.
c) To enhance the quality controlling system of Hong Kong 1:1,000 i-Series Topo Map with proposed methods and principles.

1.4 Structure of the Thesis

This thesis composes of six chapters. The first chapter is an introduction where the principles of spatial data quality are reviewed. In chapter 2, the quality evaluation process flow mentioned in ISO 19114 is elaborated. The principles of spatial data quality assessment and quality control are analyzed which includes quality assessment methods, sampling methods and quality reporting methods that mentioned in several positional accuracy standards.

Chapter 3 is the most important part in this thesis. It proposes a quality assessment procedure for topographic data which includes the principles and methods for selecting appropriate quality assessment methods and sampling strategy.

Chapter 4 conducts a case study to review the Hong Kong 1:1,000 i-Series Topo map. The production specification, data quality control and assurance measures, and data capture standards are reviewed.

Chapter 5 proposed a quality control system for Hong Kong 1:1,000 i-Series Topo map.

The last chapter is the conclusions of this study and the future developments of quality assessment of spatial data.

CHAPTER 2

AN ANALYSIS OF QUALITY ASSESSMENT AND CONTROL OF TOPOGRAPHIC DATA

Assessing quality of topographic data requires clearly defined procedures to obtain consistent results. National map agencies and International Organization for Standardization have recognized the growing needs for quality evaluation, and thus released difference types of standards and procedures, such as ASPRS Accuracy Standards for Large-Scale Maps (ASPRS, 1990), National Standards for Spatial Data Accuracy (FGDC-GPAS, 1998), and ISO 19114 (2003), etc. Establishment of methods, standards and procedures are critical step in implementing quality assessment. These standards laid the statistical foundations for other publications that determine the errors of spatial data using sampling methods.

This chapter reviews the principles of quality assessment and control of topographic data. The first part elaborates the quality evaluation procedures mentioned in ISO 19114. The second part compares seven positional accuracy standards and outlines the key points to one another. The next part reviews the sample design of quality assessment of spatial data. In the fourth part, the testing methods for positional accuracy, completeness, thematic accuracy, logical accuracy and thematic accuracy, are described. The fifth part reviews the quality reporting methods used by three digital topographic dataset and the final part gives an overview on quality control of topographic data.

2.1 Procedure for Evaluating Spatial Data Quality (ISO 19114)

ISO 19114 provides a procedure for evaluating and reporting data quality result. This section reviews the evaluation process.

2.1.1 Process Flow for Evaluating Data Quality

The process flow, proposed by ISO 19114, can be used in different stages of data product lifecycle of spatial data. The data product lifecycle includes: development of product specifications, data production, data delivery, data use and update. When developing a product specification, quality evaluation procedure is able to establish a conformance level which the final product should be met. At the production stage, it can be used for quality control. At data delivery stage, it can be used to determine data quality results. The results are used as reference information to evaluate whether a dataset conforms to its product specifications. For data use, the evaluation procedure is used to establish the conformance level with respect to users' requirements. The quality evaluation process can also use to detect the impacts on data updating. The process flow for data quality evaluation is illustrated in Figure 2.1, and the descriptions of each step are shown in Table 2.1.

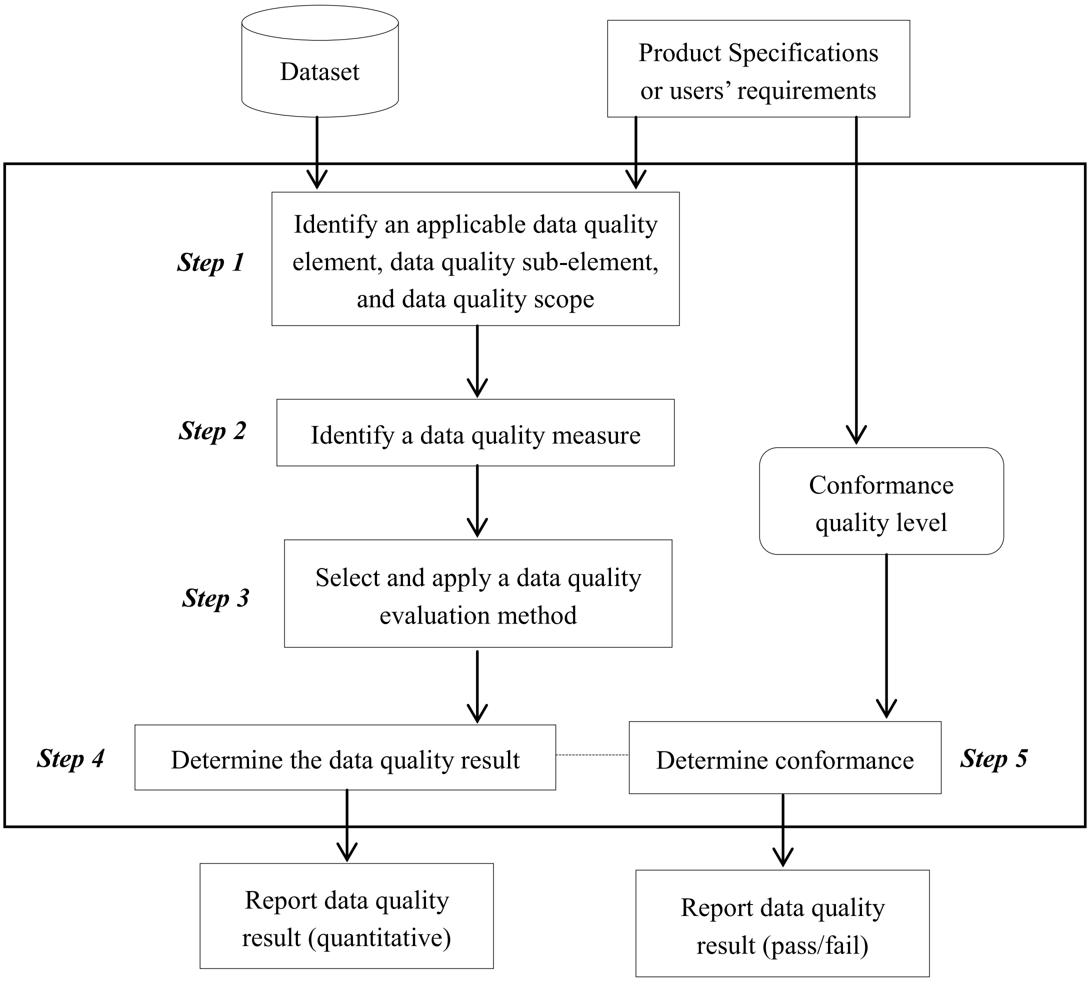


Figure 2.1 Process flow for evaluating and reporting data quality result

Table 2.1 Descriptions for each process step

Process Step	Description
1	The data quality elements, data quality sub-elements, and data quality scope to be tested are identified. This is repeated for as different tests.
2	A data quality measure is identified for each test to be performed. For example, the data quality value type and data quality value of positional accuracy can be "number" or "1.70 m" respectively.
3	A data quality evaluation method for each identified data quality measured is identified.
4	A quantitative data quality result is output of the applied method.
5	The data quality results are determined by compared with the users' requirements or conformance level defined by data provider.

2.1.2 Classification of Data Quality Evaluation Methods

Data quality evaluation methods are classified into two types: a) direct and b) indirect. Direct evaluation method assesses data quality by comparing the test data with reference information. Indirect evaluation method infers or estimates data quality using information of data production methods, data process methods and data sources. Figure 2.2 shows the classification of data evaluation methods.

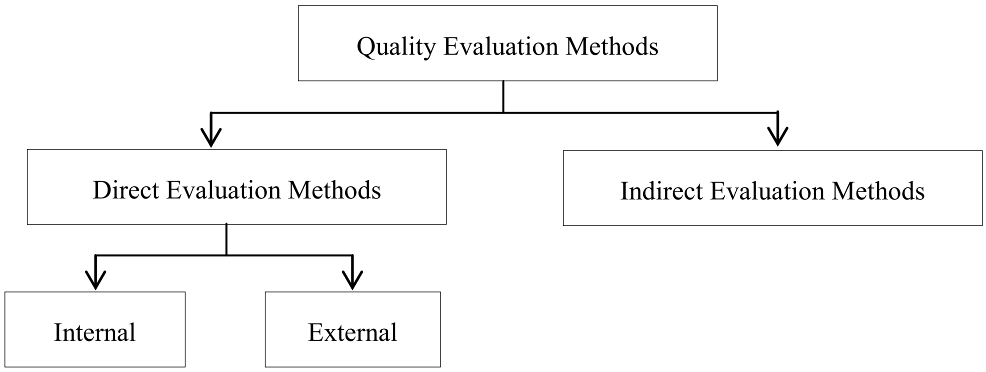


Figure 2.2 Classification of data quality evaluation methods

Direct evaluation method can be further classified into internal and external. For internal evaluation, the data needed to preform quality evaluation are internal to the dataset being evaluated, for example, silver polygon check and topology check. For external evaluation, reference data are required to the dataset being tested, for example, the positional accuracy is determined by comparing the coordinates of points of test data with reference data.

Indirect evaluation evaluates data quality based on external knowledge. The external knowledge includes product specifications, lineage, purpose, and data sources, etc.

2.1.3 Reporting Data Quality Results

ISO 19114 suggests two ways, metadata and quality evaluation report, to report data quality results. ISO 19115 has defined the metadata standards, which include data model and data dictionary, to report quantitative quality results.

When data quality results reported as metadata are only pass/fail, or aggregated data quality results are generated, it is suggested using quality evaluation report to report data quality results. The quality evaluation report is able to provide more details about the quality results and the procedures of quality evaluation. For instance, three tests have been conducted which are positional accuracy, logical consistency and completeness. The aggregated data quality result combines the results of each test and determines the acceptance of the dataset. Thus, a dataset may be deemed to be accepted even one or more individual data quality results are failure.

2.2 Positional Accuracy Standards

Positional accuracy is the most significant element for reporting quality of spatial data. According to the Glossary of the Mapping Sciences (ASPRS and ASCE, 1994), the positional accuracy is:

> "…the degree of compliance with which the coordinates of points determined from a map agree with the coordinates determined by survey or other independent means accepted as accurate."

This section reviews seven positional accuracy standards. The original documents and their amendment histories can be downloaded from official websites. The major objective of these standards is to provide methods and mathematical foundations for assessing and reporting positional accuracy. Table 2.2 gives a comparison of seven positional accuracy standards. In chronological order, these standards are as follows:

a) United States National Map Accuracy Standards (NMAS, 1947)
b) Principles of Error Theory and Cartographic Applications (Greenwalt C.R. and Schultz M.E., 1962 and 1968)
c) ASPRS Accuracy Standards for Large-Scale Maps (ASPRS, 1990)

d) The Federal Geographic Data Committee's Geospatial Positional Accuracy Standards (FGDC-GPAS, 1998)

e) ASPRS Guidelines: Vertical Accuracy Reporting for LiDAR data (ASPRS, 2004)

f) Australian Map and Spatial Data Horizontal Accuracy Standard (ICSM, 2009)

g) ASPRS Accuracy Standards for Digital Geospatial Data – Draft (ASPRS, 2014)

Table 2.2 Comparison of positional accuracy standards

	Minimum sample size	**Provide equations for estimating error statistics**	**Require assumption that errors are normally distributed**	**Uses RMSE or percentile**	**Applicability**	**Reportin g unit**
NMAS (1947)	No	No	No	- 90^{th} percentile	- Published map	Map unit
Greenwalt C.R. and Schultz M.E. (1962, 1968)	30	- Linear Error - Circular Error - Spherical Error	Yes	No	- Charting product - weapon system accuracy evaluation	Unstated
ASPRS (1990)	20	- Horizontal Accuracy - Vertical Accuracy	Not Required	No	- small-scale and large-scale map - Digital	Ground unit

					spatial data	
FGDC-GPAS (1998)	20	- Horizontal Accuracy - Vertical Accuracy	Yes	- RMSE (95% C.L.)	- Fully georeferenced map - Digital geospatial data (raster, point, vector format	Ground unit
ASPRS (2004)	20 (30 preferred) If five major land cover categories are determined, then a minimum of 100 checkpoints are required	- Vertical Accuracy	- For open terrain, normally distributed error required. - For ground categories that outside open terrain, normally distributed error not required	- RMSE (95% C.L.) - 95th percentile	- Elevation data generated using LiDAR technology	Ground unit
ICSM (2009)	20	- Horizontal Accuracy	Yes	- RMSE (95% C.L.)	- Published maps - Spatial data	Ground unit

					for geographic and geoscientific applications	
ASPRS (2014)	- Depends on project area. For example, if the project area is less than 500 km², 20 check points for horizontal check, and 20 check points for vertical check in non-vegetated area	- Horizontal Accuracy - Vertical Accuracy	- When errors are normally distributed, accuracy test is performed with RMSE value - When errors are not normally distributed and the number of check points is sufficient, accuracy test is using 95th percentile errors	- RMSE (95% C.L.) (horizontal accuracy and vertical accuracy for non-vegetated terrain) - 95^{th} percentile (vertical accuracy for vegetated terrain)	- Planimetric map - Orthophotos - Vertical accuracy for mobile mapping system, unmanned aerial system, airborne or satellite stereo imagery, LiDAR or IFSAR	Ground unit

The evolution of positional accuracy standards shows that:

a) The reporting unit is changed from map unit to ground unit.
b) Horizontal accuracy is reported at 95% C.L. using RMSE.
c) Vertical accuracy is reported at 95% C.L. using RMSE in non-vegetated terrain, and at 95% percentile in vegetated terrain.
d) No. of check points is 20 for area that less than 500 km^2.
e) Both horizontal and vertical accuracy are recommended testing with independent sources of higher accuracy.
f) The applicable products are change from paper map to digital map, orthophoto, data obtained from mobile mapping system, unmanned aerial system, satellite stereo imagery, LiDAR and IFSAR.
g) The quality assessment method for line features has been introduced.
h) The statistical analysis for both normally distributed error and not normally distributed error are introduced.

2.2.1 United States National Map Accuracy Standards (NMAS, 1947)

This standard was established to facilitate the data sharing among mapmaking agencies in U.S. It defined the horizontal accuracy and vertical accuracy, accuracy testing methods and accuracy reporting methods of manuscript and published map that of size being 15 minutes of latitude and longitude, or 7.5 minutes.

NMAS (1947) assesses and reports horizontal and vertical accuracy using map scale and contour interval. For vertical accuracy check, the apparent vertical error can be offset by permissible horizontal error. The definition of horizontal accuracy, vertical accuracy, testing method and reporting method are quoted below:

Horizontal Accuracy:

> "For maps on publication scales larger than 1:20,000, not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for map on publication scales of 1:20,000 or smaller, 1/15 inch. These limits of accuracy shall apply in all cases to positions of well-defined point only. Well-defined points are those that are easily visible or recoverable on the ground, such as monuments or makers….. In general what is well defined will be determined by what is plottable on the scale of map within 1/100 inch….."

Vertical Accuracy:

> "….. applied to contour maps on all publication scales, shall be such that not more than 10 percent of the elevations tested shall be in error more than one-half the contour interval. In checking elevations taken from the map, the apparent vertical error may be decreased by assuming a horizontal displacement within the permissible horizontal error for a map of that scale."

Accuracy Testing Method:

> "The accuracy of any map may be tested by comparing the positions of points whose locations or elevations are shown upon it with corresponding positions as determined by surveys of a higher accuracy."

Accuracy Reporting Method:

> Published map meeting these accuracy requirements shall note this fact on their legend, as "This map complies with National Map accuracy Standards".

This standard simply describes horizontal and vertical accuracy using "90th percentile", and requires well-defined points for quality assessment.

2.2.2 Principles of Error Theory and Cartographic Applications (Greenwalt C.R. and Schultz M.E., 1962 and 1968)

Greenwalt and Schultz presented theories and procedures for providing a meaningful error statement to describe the positional accuracy of spatial data. This report utilizes probability theory to calculate one-dimensional (Linear) errors, two-dimensional (Elliptical, Circular) errors, and three-dimensional (Ellipsoidal, Spherical) errors based on assumptions that errors are normally distributed. This implies that systematic errors and blunder errors have been eliminated or reduced and only random errors are remained.

The proposed principles do not stipulate a maximum error that similar to NMAS (1947). It determines the probable maximum error interval around the mean error from the sample data, and shows how to estimate the distribution of errors under different probability levels. It also provides table for converting probability level from one to another.

According to the principle, positional error should be expressed by precision indexes which are the form and probability represented by a given error. For example, let the circular probable error (CPE) of a point feature equal to 50 feet. Then, the form is circular, the magnitude is 50 feet and the probability (50 % by definition of CPE) are derived from a statistical treatment of known or estimated error components comprising the total positional error. The statement means a 50% chance that the geodetic position in question does not vary more than 50 feet from its true position. The definitions of one-dimensional error and two-dimensional error are shown below:

One-dimensional Error:

Assuming that the values of random variable $\underline{x}$ are normally distributed about the mean μ, the area under the normal probability density curve (figure 2.3) represents the total probability of the occurrence of continuous random variable $\underline{x}$ is equal to 1, or 100 %. The mathematical expression of the curve is the normal probability density function, $p(\underline{x})$:

$$p(\underline{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\underline{x}-\mu)^2}{2\sigma^2}} \qquad (2.1)$$

where: $\underline{x}$ is the random variable

μ is a parameter representing the mean value of $\underline{x}$

σ is a parameter the standard deviation, a measure or the dispersion of the random variable from the mean μ

$\sqrt{2\pi}$ is 2.5066……

e is the base of natural logarithms, 2.71828…..

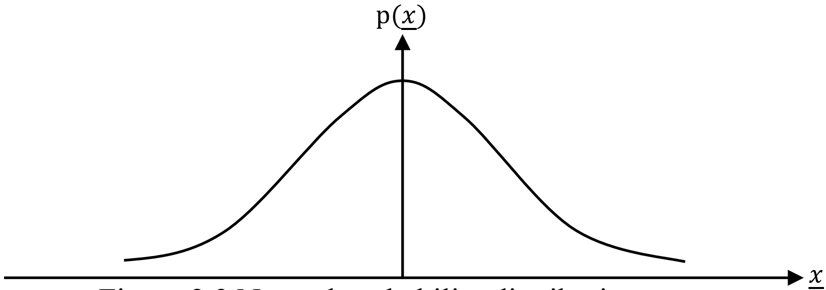


Figure 2.3 Normal probability distribution curve

Equation (2.1) expresses the normal probability density curve of an infinite number of measurements of the unknown quantity $\underline{x}$. Since the true value cannot be measured and an infinite number of measurements are impractical, estimated values obtained from a finite number or sample measurements (often, 30 values provide an adequate estimate) must be substituted for the true value and the parameters of the density function. The most probable value ($\bar{X}$) approximates the true value and is determined from the arithmetic mean of observed values:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (2.2)$$

The true error is approximated by the residual x which is the difference between the observed value and the most probable value:

$$x = X_i - \bar{X} \quad (2.3)$$

The standard error computed from a sample (σ_x) is computed from:

$$\sigma_x = \sqrt{\frac{\sum x^2}{n-1}} \quad (2.4)$$

The normal probability density function of errors now becomes

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x{}^2}} \tag{2.5}$$

A precision index reveals how errors are dispersed about zero and reflects the limiting magnitude of error for various probabilities. The standard error (σ_x) and average error (η) are two important precision index, other common usage has included three additional probability level which are precision indexes: a) probable error (PE), b) map accuracy standard (MAS) and c) the three sigma error (3σ).

The standard error (σ_x) is the most important of the indexes and has the probability of:

$$P(x) = \int_{-\sigma_x}^{+\sigma_x} p(x)dx = 0.6827 \tag{2.6}$$

Or, 68.27% of all errors will occur within the limits of $\pm\sigma_x$

The average error (η) is defines as the mean of the sum of the absolute values of all errors:

$$\eta = \frac{\Sigma|x|}{n} \tag{2.7}$$

The probability represented by the average error is 0.5751. The average error is easily computed from the standard error:

$$\eta = 0.7979\sigma_x \tag{2.8}$$

The probable error (PE) is that error which 50% of all errors in a linear distribution will not exceed. Expressed mathematically:

$$PE = \int_a^b p(x)dx = 0.50 \tag{2.9}$$

The PE can be computed from the standard error

$$PE = 0.6745\ \sigma_x \tag{2.10}$$

The U.S. National Map Accuracy Standards (NMAS, 1947) specify that no more than 10% of map elevations (one-dimensional error) shall be in error by more than a given limit. This can be interpreted as limiting the size of error of which 90% of the elevations will not exceed. Therefore, the map accuracy standard (MAS) is represented by:

$$PE = \int_a^b p(x)dx = 0.90 \tag{2.11}$$

or, computed from the standard error:

$$MAS = 1.6449\ \sigma_x \tag{2.12}$$

The three sigma error is an error three times of the magnitude of the standard error. The probability represented by 3σ error is 99.73%. Figure 2.4 shows the probability areas of the standard normal distribution corresponding to the probability levels of PE, σ_x, MAS and 3σ.

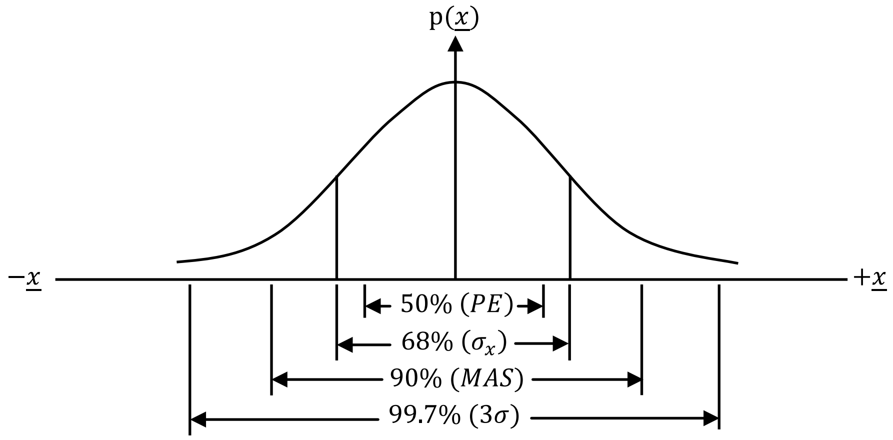


Figure 2.4 Probability areas of PE, σ_x, MAS and 3σ

Two-dimensional Error:

A two-dimensional error can be applied to error of a geographic position of a point referred to X and Y axes. Each observation of the X and Y coordinates will contain errors "x", "y". When assumed random and independent, each error has a probability density distribution of:

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_x{}^2}}$$

And;

$$p(y) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_y{}^2}}$$

Since x and y are two random and independent events, the probability of them occurring simultaneously is equal to the product of their individual probabilities. The two-dimensional density function becomes:

$$p(x,y) = p(x) \cdot p(y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x{}^2} + \frac{y^2}{\sigma_y{}^2}\right)} \ldots\ldots\ldots\ldots\ldots\ldots(2.13)$$

The probability density function integrated over a certain region becomes the probability distribution function which yields the probability that x and y will occur simultaneously within that region:

$$P(x,y) = \int\int p(x,y)dxdy$$

Since both positive and negative value of either "x" and "y" will occur with equal frequency, the errors may be considered as radial errors, designated by "r", where

$$r = \sqrt{x^2 + y^2}$$

The probability distribution function of the radial error expressing the probability that "r" will be equal to or less than radius R, or the probability that the vector xy will be contained within a circle of radius R, is stated as:

$$\mathrm{P(R)} = \frac{1}{\sigma_x\sigma_y}\int_0^R r e^{-\frac{r^2}{4\sigma_y{}^2}\left[1-\frac{\sigma_y{}^2}{\sigma_x{}^2}\right]} I_o\left[\frac{r^2}{4\sigma_y{}^2}\left(\frac{\sigma_y{}^2}{\sigma_x{}^2} - 1\right)\right] dr \ldots\ldots\ldots\ldots\ldots(2.14)$$

A special case the P(R) function (2.14) is formed when r = R, and $\sigma_x = \sigma_y = \sigma_r$

$$P(R) = P_c = 1 - e^{-\frac{R^2}{2\sigma_c^2}} \quad \text{.............................(2.15)}$$

where: P_c is the circular probability distribution function, a special case of P(R)

R is the radius of the probability circle

σ_c is the circular standard error, a special case of σ_r when $\sigma_x = \sigma_y = \sigma_r$

When σ_x *and* σ_y are not equal, the P(R) function (2.14), is modified by letting "a" equal to the ratio $\frac{\sigma_x}{\sigma_y}$ where σ_x is the smaller standard error of two. P(R) now becomes:

$$P(R) = \frac{2a}{1+a^2} \int_0^x e^{-v} I_0(vk)\,dv \quad \text{..........................(2.16)}$$

where: $x = \frac{R^2}{4\sigma_y^2}\left[\frac{1+a^2}{a^2}\right]$

$v = \frac{r^2}{4\sigma_y^2}\left[\frac{1+a^2}{a^2}\right]$

$k = (\frac{1-a^2}{1+a^2})$

The precision indexes illustrated in figure 2.5 are measures of the dispersion of errors in a distribution and represent the error which is unlikely to be exceeded for a given probability. The preferred circular precision indexes, consistent with indexes used in the linear distribution are: a) the circular standard error (σ_c), b) the circular probable error (CPE), c) the circular map accuracy standard (CMAS) and d) three-five sigma ($3.5\sigma_c$).

The probability of the circular standard error is found by solving the equation (2.5) for P_c where $\sigma_r = R$, thus

$$P_c = 1 - e^{-\frac{\sigma_c{}^2}{2\sigma_c{}^2}}$$

$$= 1 - e^{-\frac{1}{2}}$$

$$= 1 - 0.60653 = 0.3935 \qquad (2.17)$$

That is, 39.35% of all errors in a circular distribution are not expected to exceed the circular standard error.

When σ_x and σ_y are not equal, a normal circular error distribution may be substituted for the elliptical distribution. The substitution is satisfactory for the error analysis within specified $\sigma_{min}/\sigma_{max}$ ratio (where σ_{min} is the minimum linear standard error of the two) ratio. For the $\sigma_{min}/\sigma_{max}$ ratio between 1.0 and 0.6, the curve is a straight line with the equations:

$$\sigma_c \sim (0.5222\,\sigma_{min} + 0.4778\,\sigma_{max}) \qquad (2.18)$$

A rapid approximation gives a slightly larger σ_c value for the same $\sigma_{min}/\sigma_{max}$ ratio:

$$\sigma_c \sim 0.5(\sigma_x + \sigma_y) \qquad (2.19)$$

The circular probable error (CPE) is the circular error which 50% of all errors in a circular distribution will not exceed, or the value of R in equation (2.15) which makes $P_c = 0.5$. The CPE in a truly circular distribution (i.e. $\sigma_x = \sigma_y = \sigma_c$) is:

$$CPE = 1.774\sigma_c \quad (2.20)$$

When $\sigma_x and\ \sigma_y$ are not equal, an approximate CPE is determined from equation (2.16).

$$\text{CPE} \sim (0.6142\ \sigma_{min} + 0.5632\ \sigma_{max}) \quad (2.21)$$

where $\sigma_{min}/\sigma_{max}$ ratio is between 1.0 and 0.3

$$\text{CPE} \sim (0.4263\ \sigma_{min} + 0.6196\ \sigma_{max}) \quad (2.22)$$

where $\sigma_{min}/\sigma_{max}$ ratio is between 0.3 and 0.2

A rapid approximation of the CPE plots as a straight line which intersects the 50% probability curve at the point where $\sigma_{min}/\sigma_{max}$=2 and has the equation:

$$\text{CPE} \sim 0.5887(\ \sigma_x +\ \sigma_y) \quad (2.23)$$

when $\sigma_{min}/\sigma_{max}$ is between 1.0 and 0.2

The mean square positional error (MSPE) is defined as the radius of the error circle equal to $1.4142\sigma_c$. When $\sigma_x and\ \sigma_y$ are approximately equal, the MSPE defines the error in a geographic position and is computed as:

$$\text{MSPE} = \sqrt{{\sigma_x}^2 + {\sigma_y}^2} \quad (2.24)$$

where $\sigma_{min}/\sigma_{max}$ ratio is between 0.1 and 0.8

The circular map accuracy standard is based on the percentage level in the use by U.S. National Map Accuracy Standard (NMAS, 1947) when specify that no more than 10% of the well-defined points in a map will exceed a given error. The circular map accuracy standard is represented by the value of R in equation (2.15) when $P_c = 0.90$ and is computed:

$$CMAS = 2.1460\sigma_c \quad (2.25)$$

$$\text{or } CMAS = 1.8227\ CPE \quad (2.26)$$

The three-five sigma error, representing a circular probability of 99.78% in a circular distribution and has a magnitude 3.5 time that of the circular standard error.

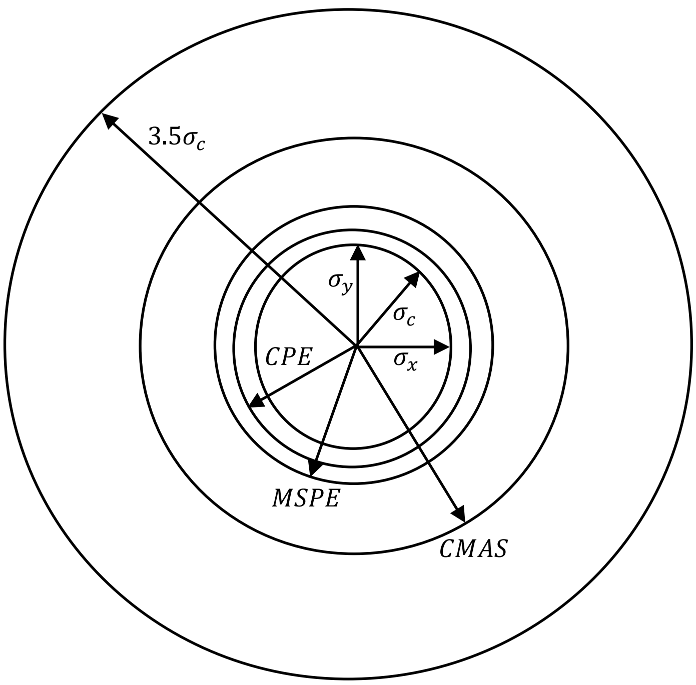


Figure 2.5 Normal circular distributions

2.2.3 ASPRS Accuracy Standards for Large-Scale Maps (ASPRS, 1990)

Similar to the NMAS (1947), ASPRS (1990) stipulates conformance levels of large-scale map products. The positional accuracy of digital spatial data can be indicated at ground unit that related to the appropriate map scale for graphic presentation. In this standard, maps are classified into Class 1, Class 2, and Class 3 according to their accuracy. It also establishes threshold (RMSE) for both horizontal accuracy and vertical accuracy.

The definitions of horizontal accuracy, vertical accuracy and accuracy testing methods are quoted below:

Horizontal Accuracy:

> "Horizontal map accuracy is defined as the root-mean-square-error (RMSE) in terms of the x, y coordinates for checked points as determined at ground scale of the map. The RMSE is the cumulative result of all errors including those introduced by the processes of ground surveys, map compilation and final extraction of ground dimensions from the map."

The standard establishes the limiting RMSE for map that with typical map scales. The accuracy tests are applied to well-defined that can be sharply identified as discrete points in both test data and reference data. The limiting RMSE for Class 1 map is shown in Table 2.3. Maps compile within the limiting RMSE of twice or three times those allowed for Class 1 map are designated as Class 2 or Class 3 maps respectively.

Table 2.3 Planimetric coordinate accuracy requirement for well-defined points – Class 1 Maps

(Indicates the practical limit for aerial methods – for scales above this line, ground methods are normally used)

Limiting RMSE (meters)	Typical Map Scale
0.0125	1:50
0.025	1:100
0.050	1:200
0.125	1:500
0.25	1:1,000
0.50	1:2,000
1.00	1:4,000
1.25	1:5,000
2.50	1:10,000
5.00	1:20,000

Vertical Accuracy:

> "Vertical map accuracy is defined as the RMSE in evaluation in terms of project's evaluation datum for well-defined point only. For Class 1 map the limiting RMSE in evaluation is set by the standard at one-third the indicated contour interval for well-defined points only. Spot height shall be shown on the map within a limiting RMSE of one-sixth of the contour interval."

Accuracy Testing Method:

> "Testing for horizontal accuracy compliance is done by comparing the planimetric coordinates of well-defined ground points to the coordinates of the same points as determined by a horizontal check survey of higher accuracy. The check survey shall be designed according to the Federal Geodetic Control Committee (FGCC, 1984) standards and specifications to achieve standard deviations equal to or less than one-third of the limiting RMSE select for the map."

> "Testing for vertical accuracy compliance shall be accomplished by comparing the elevations of well-defined points as determined from the map to corresponding elevations determined by a survey of higher accuracy. For purposes of checking elevations, the map position of the ground point may be shifted in any direction by an amount equal to twice the limiting RMSE in position. The vertical check survey should be designed to produce RMSE in elevation differences at check point location no larger than $1/20^{th}$ of the contour interval."

This standard recommends that the check points should be distributed more density near important features and sparsely in other areas. For map sheet that has standard dimensions, it is intended to portray a uniform spatial accuracy over the entire map sheet and the error represented shall be reasonable distributed. For example, if a test uses a minimum of twenty check points, it is suggested that at least 25% of the points be located in each quadrant of the map sheet and these points are spaced at intervals equal to at least 10% of the map sheet diagonal.

2.2.4 The Federal Geographic Data Committee's Geospatial Positional Accuracy Standards (FGDC-GPAS, 1998)

The Federal Geographic Data Committee established guidelines for measuring, analyzing, and reporting positional accuracy of spatial data that derived from various data source such as aerial photograph, satellite imagery, and ground survey, etc. This standard is widely accepted by national mapping agencies and private sectors.

The FGDC-GPAS (1998) explicitly rejects setting a threshold accuracy value. Rather, it encourages that the threshold should be determined as needed. It suggests reporting positional accuracy in "ground distance at 95% confidence level" using RMSE.

The FGDC-GPAS (1998) comprises of five parts:

Part 1: Reporting Methodology

Part 2: Standards for Geodetic Networks

Part 3: National Standard for Spatial Data Accuracy (NSSDA, 1998)

Part 4: Standards for Architecture, Engineering, Construction (A/E/C) and Facility Management

Part 5: Standards for Nautical Charting Hydrographic Surveys – Public Review Draft

Parts 1 to Part 3 are relevant to positional accuracy assessment, so they are reviewed in the following three sections.

2.2.4.1 Part 1: Reporting Methodology

This part is an introduction. The FGDC-GPAS (1998) intends to provide a method for reporting horizontal and vertical accuracy of clearly defined point features. It compares the coordinate values of testing points (e.g. a cartographically-derived value) with the coordinates of reference points (e.g. Global Positioning System (GPS) geodetic network survey) in independent dataset.

2.2.4.2 Part 2: Standards for Geodetic Networks

This part provides a method for determining and reporting the horizontal and vertical accuracy of geodetic networks. Accuracy of geodetic networks is classified in accordance with their accuracy level. This standard states a procedure to classify the accuracy of control network and the accuracy reporting methods. The accuracy reporting method is quoted below.

Accuracy Reporting:

> “When providing geodetic point coordinate data, a statement should be provided that the data meets a particular accuracy standard for both the local accuracy and the network accuracy. For example, these geodetic control data meet the 2-centimeter local accuracy standard for the horizontal coordinate values and the 5-centimeter local accuracy standard for the vertical coordinate values (heights) at 95-percent confidence level……….”

2.2.4.3 Part 3: National Standard for Spatial Data Accuracy (NSSDA, 1998)

This part provides methodology for estimating the positional accuracy of points in spatial data. This standard recommends that users shall identify acceptable accuracies for their products or applications, while data providers must determine the accuracy of their data and report it according to NSSDA.

Accuracy Test Guidelines:

> "According to the Spatial Data Transfer Standard (SDTS, 1998), accuracy testing by an independent source of higher accuracy is the preferred test for positional accuracy………
> A minimum of 20 check points shall be tested, distributed to reflect the geographic area of interest and the distribution of error in the dataset. When 20 points are tested, 95% confidence level allows one point to fail the threshold given in product specifications. If fewer than twenty points can be identified for testing, use an alternative mean to evaluate the accuracy of the dataset such as "Deductive Estimate", "Internal Evidence" and "Comparison to Source."

Accuracy Reporting:

> "Positional accuracy values shall be reported in ground distance……... Accuracy reporting in ground distance allows users to directly compare dataset of differing scales or resolution……… A simple statement of conformance is not adequate in itself. Measures based on map characteristics, such as publication scale or contour interval,

> are no longer adequate when data can be readily manipulated and output to any scale or to different data formats."

2.2.5 ASPRS Guidelines: Vertical Accuracy Reporting for LiDAR Data (ASPRS, 2004)

Vertical accuracy is important in specifying the quality of elevation data. This standard aims to provide guidelines for testing and reporting elevation data generated by Light Detection and Ranging (LiDAR) or related technology. When testing the elevation data derived from LiDAR data, it is important to specify the vertical accuracy determined for final product. For instance, when a digital elevation model (DEM) or digital surface model (DSM) is generated from LiDAR mass points, a TIN may first be generated before the DEM or DSM are derived. The accuracy shall not be tested for the TIN because it may contain greater error, especially when generalization or surface smoothing has been applied to the final product.

Vertical Accuracy:

> "With the NSSDA (1998) the vertical accuracy of a dataset ($Accuracy_z$) is defined by the RMSE of the elevation data in terms of feet or meters at ground scale, rather than in terms of the published map's contour interval. Because the NSSDA (1998) does not address the suitability of data for any particular product, map scale, contour interval, or other application, no error threshold are established by the standard. However, it is often helpful to use familiar NMAS (1947) thresholds for determining reasonable NSSDA (1998)

accuracy requirements for various types of terrain and relief."

This relationship is:

$$NMAS\ CI = 3.2898 \times\ RMSE_z \tag{2.2.5.1}$$

$$NMAS\ CI =\ Accuracy_z\ /\ 0.5985 \tag{2.2.5.2}$$

Where:

$$Accuracy_z = 1.9600 \times\ RMSE_z\ (Normally\ distributed\ error) \tag{2.2.5.3}$$

When the error is not normally distributed, this standard recommends $Accuracy_z$ be determined by 95th percentile testing, instead of using equation 2.2.5.3. In statistical test, normal distribution can be tested for by calculating the skewness of testing data. If the skew exceeds ± 0.5, this indicates an asymmetry in the data and further investigation should be conducted.

ASPRS also suggests that the vertical accuracy of LiDAR derived elevation data shall be reported according to different ground cover categories, such as open terrain, tall weeds, brush lands, forested areas fully covered by trees or urban, etc.

Horizontal Accuracy:

Since elevation data, DEM and DSM, often lacks of well-defined points, ASPRS recommends reporting the expected horizontal accuracy of elevation data as determined by system studies such as orientation determination, errors in INS and GPS, coordinate transformation, errors in ground control points, and errors in laser sensor, etc.

2.2.6 Australian Map and Spatial Data Horizontal Accuracy Standard (ICSM, 2009)

This standard provides guidelines for calculating and reporting absolute horizontal accuracy of spatial data. The horizontal accuracy is reported at 95% confidence level and in ground distance using metric units. ICSM (2009) recommends using three ways to determine horizontal accuracy:

a) Testing against an independent source of higher accuracy (similar to NSSDA, 1998)
b) Testing by deductive estimate
c) Testing by inference

Accuracy of Linear Features:

ICSM (2009) provides a method to test the horizontal accuracy of linear feature. This standard suggests using equally spaced perpendicular offset along the tested feature to their intersection with the independent source of higher accuracy to test the positional accuracy of linear features in dataset. This method is limited to regular bend features only, because it may not be logically matched the locations on tested features with their corresponding location on an independent source. In these cases, the offset to the independent source should be adjusted so that a logical matching is achieved.

Testing by Deductive Estimate:

This method estimates the propagation of errors through entire data production process, such as data capture, data conversion, data editing and manipulation, etc. The errors can be estimated using following equation:

$$95\%Accuracy = 1.96 \times S = 1.96\sqrt{{S_m}^2 + {S_l}^2 + {S_{man}}^2}$$

Where:

S is the standard deviation and the subscripts m, l and man refer to source materials, conversion process and manipulation processes respectively.

Testing by Inference:

This method assumes that maps go through same production process and with same scale would be reasonable to have same accuracy level. If the accuracy of one map in a map series is tested, the accuracy of all subsequent maps in the series can be inferred.

2.2.7 ASPRS Accuracy Standards for Digital Geospatial Data – Draft (ASPRS, 2014)

This standard aims to replace the ASPRS (1990) and ASPRS (2004) with new accuracy standards that better address digital orthophoto and digital elevation data. It utilizes Ground Sampling Distance (GSD) to report the accuracy of digital imagery. The GSD is the linear dimension of a sample pixel's footprint on the ground in the source image, and it is assumed that "pixel size" is the real-world's ground size of a pixel in a digital orthophoto product after all rectifications and resampling procedures have been performed. Furthermore, GSD is intended to pertain to near-vertical imagery and not to oblique imagery, also recognizing that GSD values can vary greatly in cities and mountainous areas.

Generally, the testing method of horizontal and vertical accuracy is the same as

ASPRS (1990). However, elevation dataset rarely includes well-defined point features, and it is extremely difficult and expensive to acquire surveyed vertical check points at the exact same horizontal coordinates in LiDAR mass points. To deal with this problem, Triangulated Irregular Networks (TINs) of elevation dataset are interpolated at the horizontal coordinates of vertical check points in order to interpolate elevations at those coordinates for the dataset being tested. It is suggested using high density elevation dataset so that interpolated elevation errors are minimized. When terrain is flat or has uniform slope, interpolation errors are significantly reduced. Therefore, the vertical check points should be surveyed on flat or uniformly-sloped terrain, with slopes of 10 percent or less.

Horizontal Accuracy Standard for Digital Orthophotos:

The horizontal accuracy is tested and reported using RMSE statistics. Class I product refers to high-accuracy survey-grade geospatial data for more-demanding engineering application, Class II product refers to standard, high-accuracy mapping-grade geospatial data, and Class III product refers to lower-accuracy visualization-grade geospatial data suitable for less-demanding user applications. Table 2.4 includes three standard ASPRS horizontal accuracy class (I, II, III) applicable to digital orthophoto produced from digital imagery with any GSD. It is the pixel size of the final digital orthophoto being tested and is used to establish horizontal accuracy classes for digital orthophoto.

Table 2.4 Horizontal accuracy standards for orthophotos

($RMSE_x$ and $RMSE_y$ equal the horizontal linear $RMSE$ in the easting direction and northing direction respectively. Class N refers to any accuracy class that suits the project.)

Horizontal Accuracy Class	$RMSE_x$ **and** $RMSE_y$	**Orthophoto Mosaic Seamline Maximum Mismatch**	**Aerial Triangulation or INS-based** $RMSE_x$ **,** $RMSE_y$ **and** $RMSE_z$
I	Pixel size × 1.0	Pixel size × 2.0	Pixel size × 0.5
II	Pixel size × 2.0	Pixel size × 4.0	Pixel size × 1.0
III	Pixel size × 3.0	Pixel size × 6.0	Pixel size × 1.5
….	…	…	…
N	Pixel size × N	Pixel size × N	Pixel size × 0.5N

The aerial triangulation or the INS-based sensor orientation plays a dominant role in determining the accuracy of the final mapping products. Ground control points should be at least three times better than the expected accuracy of aerial triangulation. For example, to produce a 15 cm orthophoto, the ground control to be used should have $RMSE_{xyz}$ of 2.5 cm considering the required aerial triangulation $RMSE_{xyz}$ of 7.5 cm (1/2 the orthophoto's pixel size). When producing digital orthophoto, the pixel size should never be less than 95% of the GSD of the raw imagery acquired by the sensor.

Horizontal Accuracy Standard for Planimetric Maps:

This part defines horizontal accuracy standard for different classes of planimetric maps using Map Scale Factors. Table 2.5 includes three ASPRS horizontal accuracy classes (I, II and III). The Class I accuracy formula is based on the map's scale factor, which is the reciprocal of the ratio used to specify the map scale. The derivation of the number 0.0125 is 1.25% of the Map Scale Factor. For example, if a map was compiled for use at a scale of 1: 1,200, the Scale Factor is 1,200. Then the RMSE in X or Y (cm) = 0.0125 times the Scale Factor. Referring to table 2.5, the Class I $RMSE_x$ and $RMSE_y$ standard value would be 1,200 * 0.0125 = 15 cm. The 0.0125, 0.025 and 0.037 multipliers that are applied only to RMSE value computed in centimeters.

Table 2.5 Horizontal accuracy standards for digital planimetric data

Horizontal Data Accuracy Class	$\boldsymbol{RMSE_x}$ **and** $\boldsymbol{RMSE_y}$ **(cm)**
I	1.25% of Map Scale Factor (0.0125 × Map Scale Factor)
II	2 × Map Scale Factor (0.025 × Map Scale Factor)
III	3 × Map Scale Factor (0.0375 × Map Scale Factor)
….	…
N	N × Class I Accuracy

Vertical Accuracy Standards:

The vertical accuracy of digital elevation data is suggested to report at 95% confidence level and 95th percentile for non-vegetated terrain and vegetated terrain respectively. Table 2.6 shows the vertical accuracy standards for digital elevation data. The Non-vegetated Vertical Accuracy (NVA) is estimated by multiplying the $RMSE_z$

by 1.96 and reporting at 95% confidence level. The Vegetated Vertical Accuracy (VVA) is computed at the 95th percentile of absolute errors in all types of land categories. The relative accuracy between LiDAR and IFSAR swaths in overlap area is a measure of the quality of the system calibration and bore-sighing. A dataset cannot be any more accurate than its component parts (swaths) are accurate relative to each other. The requirements for relative accuracy are therefore more rigorous than those for absolute accuracy.

Table 2.6 Vertical accuracy standards for digital elevation data

Vertical Data Accuracy Class	**$RMSE_z$ in Non-Vegetated Terrain (cm)**	**Non-Vegetated Vertical Accuracy (NVA) at 95% Confidence Level (cm)**	**Vegetated Vertical Accuracy (VVA) at 95th Percentile (cm)**	**Relative Accuracy Swath-to-Swath in Non-Vegetated Terrain ($RMSE_z$/ Max Diff) (cm)**
I	1.0	2.0	2.9	0.8/1.6
II	2.5	4.9	7.4	2.0/4.0
III	5.0	9.8	14.7	4.0/8.0
IV	10.0	19.6	29.4	8.0/16.0
V	12.5	24.5	36.8	10.0/20.0
VI	20.0	39.2	58.8	16.0/32.0
VII	33.3	65.3	98.0	26.7/53.3
VIII	66.7	130.7	196.0	53.3/106.6
IX	100.0	196.0	294.0	80.0/160.0
X	333.3	653.3	980.0	266.6/533.4

Number of Checkpoints:

This standard suggests using 100 static vertical checkpoints per each 2,500 km^2 within the project. This provides a statistically defensible number of samples for vertical accuracy assessment. Vertical checkpoints are not clearly-defined point features. It is also recommended 60 static horizontal checkpoints for each 2,500 square kilometer area within the project, while horizontal checkpoints must be clearly-defined point features, clearly visible on the digital orthophoto or planimetric maps being tested. Table 2.7 shows the recommended number of checkpoints based on project area.

Table 2.7 Recommended numbers of checkpoints based on project area

Project Area (Square Kilometers)	**Horizontal Testing**	**Vertical Testing (not clearly-defined checkpoints)**		
	Total Number of Static Horizontal Checkpoints (clearly-defined points)	**Number of Static Vertical Checkpoints in NVA**	**Number of Static Vertical Checkpoints in VVA**	**Total Number of Static Vertical Checkpoints**
≤500	20	20	0	20
501 – 750	25	20	10	30
751 – 1000	30	25	15	40
1001 – 1250	35	30	20	50
1251 – 1500	40	35	25	60
1501 – 1750	45	40	30	70

1751 – 2000	50	45	35	80
2001 – 2250	55	50	40	90
2251 - 2500	60	55	45	100

2.3 Sampling Methods

Sampling design is one of important components in spatial data quality assessment, because this determines both the cost and the rigor of the assessment. Assessing the quality of topographic dataset requires an efficient and effective sampling method to collect reference data. This section reviews the sampling strategy which corresponding to spatial sampling techniques to geographic data.

ISO 19114 provides different sampling methods for assessing spatial data. It categorizes sampling methods into two aspects, which are: feature or area, and probability or judgment. Figure 2.6 illustrates the sampling strategy relationships.

Feature-guided sampling collects samples based on non-spatial attributes of the feature instead of their spatial location. This sampling method selects features randomly using their attribute information, such as building name and address, and assuming homogeneous production characteristic for the entire quality of dataset. Area-guided sampling collects samples based on location considerations. The sampling units can be graphical extents, such as map sheet and gridded data.

Probabilistic sampling uses sampling theory and selects sample features randomly from the dataset. In probabilistic sampling, statistical inferences can be made about the sampled population such that each member of the population from which the sample selected has a known probability of selection. Judgmental sampling selects features based on expert knowledge or professional experiments.

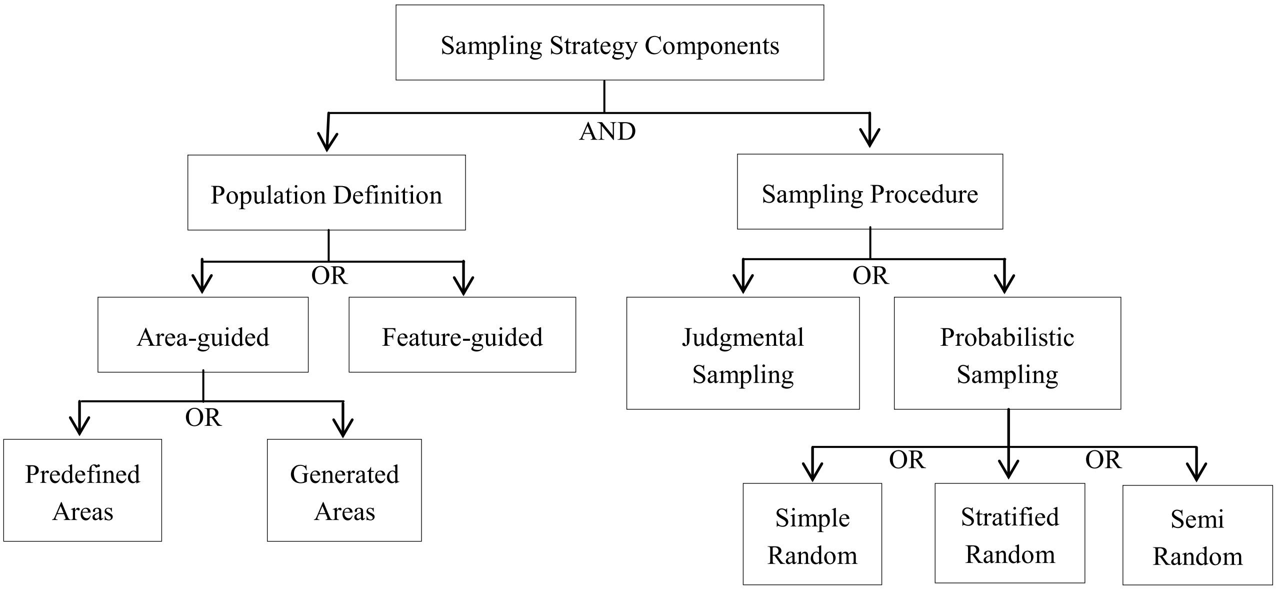


Figure 2.6 Sampling strategy relationships

W.G. Cochran (1977), in his book Sampling Techniques, has proposed eleven steps in planning and executing a sampling survey. These steps can be applied in sampling design for quality assessment of spatial data. These steps are demonstrated below:

Steps 1 - Objectives of the tests:

Quality assessment of topographic dataset may be separated into various tests based on the quality elements that have been selected. A lucid statement of the objective of each test is important. For example, "measure the error distance between absolute coordinate value of well-defined building corners in the dataset and those in the universe of discourse, and compute the RMSE from the error distance" can be an objective of a test.

Steps 2 - Population to be sampled:

The population is used to denote the aggregated from which the sample is chosen.

When sampling a population of spatial data, the samples can be selected from a group of data that shares common characteristics. Common characteristic can be treated as same data collection criteria, same original sources, or same terrain types.

Steps 3 - Data to be collected:

This step designs what kind of data to be collected for analysis. It is important to verify that all data being collected shall fulfill the requirements of each test. For instance, positional accuracy assessment requires an independent dataset that have three times more accurate than the accuracy of the testing dataset.

Steps 4 - Degree of precision desired:

The results of tests are always subject to some uncertainty because only part of population has been measured and errors occurred in measurement. This uncertainty may be reduced by taking larger samples and by using more accurate reference data, but this costs time and money. Therefore, the specification of the degree of precision desired in the results is crucial.

Steps 5 - Methods of data quality evaluation:

There are diverse methods for assessing the quality of spatial data, such as deductive estimate, comparison to source, tests based on polygon overly, etc. (SDTS, 1998). As mentioned in section 2.1.2, these methods can be classified into two types: a) direct evaluation methods, and b) indirect evaluation methods.

Steps 6 - The frame:

Before selection of the sample, the dataset must be divided into parts that are calling sampling units, or units. Sampling units or units are the area of the dataset where evaluation is conducted. These units may cover the whole dataset and they must not overlap, for examples, an area of rectangle survey sheet. The construction of these sampling units is called a frame.

Steps 7 - Define the sampling sample size:

Sample size is the information on how many features or attributes on average are extracted for evaluation. Sampling size is directly related to the costs and time involved for each test. Table 2.2 has summarized the sample size of the commonly used positional accuracy standards. Normally, these standards recommend that the minimum sample size for positional accuracy check is 20. ISO 2859 and ISO 3951 also provides sampling plan for non-spatial products. ISO 2859-1 presents a popular acceptance sampling plans using attributes which requires prohibitively large sample sizes. ISO 3951-1 presents a popular acceptance sampling plans using variables which the sample sizes are much smaller. ISO 3951-5 proposes a sequential sampling plan which has the smallest average sample size.

Steps 8 - The pretest:

It is found useful to try out the field works or quality evaluation on a small scale. This nearly always results in improvement in the quality evaluation process and may reveal other troubles that will be serious on a large scale.

Steps 9 - Organization of the quality evaluation work:

Plans must be made for handling problems occur in different types of quality evaluation methods. Field check, graphic check, and automatic program check are common ways for quality evaluation. Field check is able to evaluate completeness, positional accuracy and thematic accuracy. It obtains “true” value of the features in the real world and takes them as reference data for evaluation. Graphic check is an indoor data quality evaluation method, which includes: a) hard copy map check, b) on screen check, and c) computer-human interactive check. Automatic program check does not involve any human interference. It is executed using computer software or customized programs. Without human interference, the quality inspection algorithm greatly enhances the efficiency and accuracy of quality evaluation works.

Steps 10 - Analysis and report of the data quality result:

This step analyzes and reports the data quality result. ISO 19138 introduces the concept of data quality measures which presented two principle categories named: a) counting-related data quality measures, and b) uncertainty-related data quality measure. The counting-related data quality measures are based on the concept of counting errors or correct items. The uncertainty-related data quality measures are based on the concept of modelling the uncertainty of measurement with statistical methods. The details of analyzing and reporting data quality results are presented in section 3.3 and 3.4.

Steps 11 - Information gained for future data quality assessment:

Any executed sampling plan is potentially a guide to future sampling, in the result that it provides information about the means, standard deviations, and nature of the variability of the principal measurements and about the costs involved in obtaining the results.

2.3.1 Simple Random Sampling

This method selects samples out of the population such that every one of distinct samples has an equal chance of being drawn. Simple random sampling is useful when the testing data has relatively homogeneous in the characteristics, i.e. no major patterns and clusters. The disadvantage of using simple random sampling is that the selected samples may cluster in part of the testing area.

In simple random sampling, the first draw the probability that some one of the *n* specified units is selected is *n/N*, and at the second draw the probability that some one of the remaining (n – 1) specified units is drawn is (n – 1)/(N – 1), and so on. Hence the probability that all *n* specified units are selected in *n* draws is:

$$\frac{n}{N}\cdot\frac{n-1}{N-1}\cdot\frac{n-2}{N-2}\cdots\cdots\frac{1}{N-n+1}=\frac{n!\,(N-n)!}{N!}=\frac{1}{{}_NC_n}$$

Since the number that has been drawn is removed from the population for all subsequent draws, this method is also named random sampling without replacement.

2.3.2 Stratified Random Sampling

In this sampling method, the population is first divided into subpopulations that are non-overlapping, and together they comprise the whole of the population. The subpopulations, named strata, are more homogeneous among samples items in the same strata than among samples items in other strata. This sampling strategy has the greater precision in estimating the mean and variance than that of simple random sampling.

When the strata have been determined, samples are drawn from each and the drawings are made independently in different strata. If a simple random simple is taken in each stratum, the whole procedure is defined as stratified random sampling. This sampling method is commonly used in assessing quality of spatial data, as most spatial data are collected with different data sources and with different precisions. If each stratum is homogeneous, and the measurements vary little from one to another, a precise estimate of any stratum mean can be obtained and then combined into a precise estimate for the whole population.

2.3.3 Semi-random Sampling (Systematic Sampling)

Semi-random sampling uses random selection of the initial sample items. Systematic grid sampling is an example of semi-random sampling which is useful for estimating spatial trends or pattern. This method provides a mean to ensure the distribution and coverage of testing area.

2.3.4 Feature-guided Sampling (Non-spatial Sampling)

This sampling method relies on non-spatial attributes, such as feature code, feature type or data source. Features with homogeneous production characteristic are thus selected randomly within dataset. When the features are selected randomly in the dataset, there is a risk of the occurrence of a sample being located in a small area. Thus, semi-random sampling method can be used to verify the sample size and location of the samples selected by feature-guided sampling methods.

2.3.5 Area-guided Sampling (Spatial Sampling)

Area-guided sampling selects sample items based on location consideration. Square or rectangle grids, and political or statistical area are common units for area-guided sampling. This method usually used as a first stage of sampling, followed by a feature-guided sampling with each subarea. Figure 2.7 (a) and (b) illustrates two examples using rectangle grid as sampling unit to assess the quality of dataset. When using a regular grid cell size, a rule is needed to include or exclude cells that are not completely inside the area of testing dataset. In addition, if the distribution of features is non-homogeneous, different size sampling unit is needed in semi-random sampling.

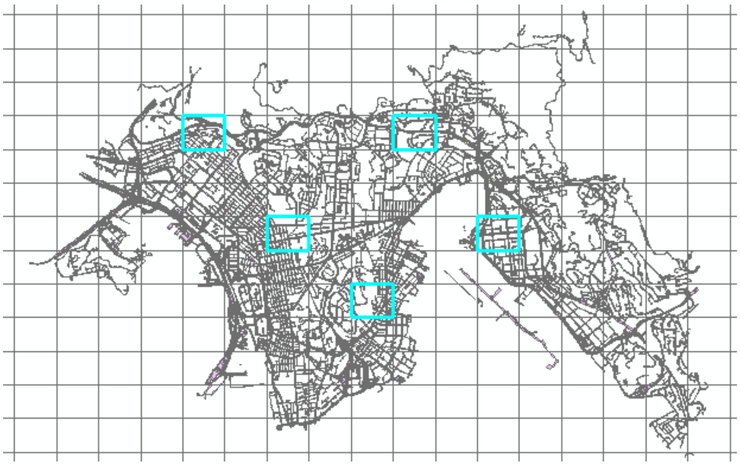


Figure 2.7 (a) Example of area-guided sampling using random selection

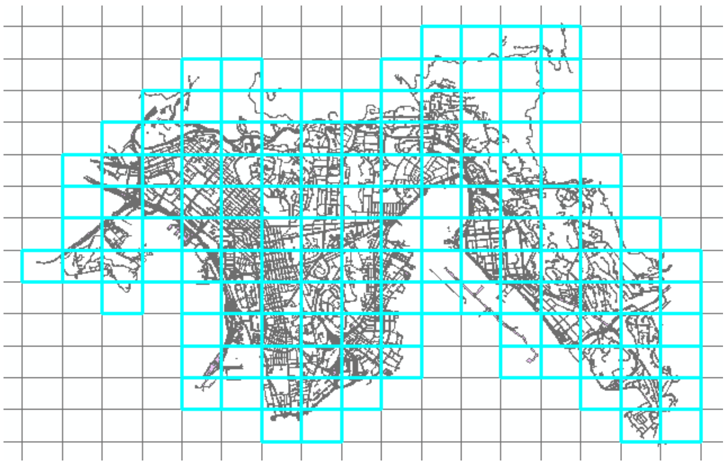


Figure 2.7 (b) Example of area-guided sampling using non-random selection

2.4 Quality Assessment Methods

SDTS (1998) introduced various methods to assess positional accuracy, attribute accuracy, logical consistency and completeness and they can be grouped into four categories, which are: test based on comparison of independent source of higher accuracy, test by deductive estimate, test by inference, and specific logical and topological tests.

When a higher accuracy reference data covers or partially covers the spatial or thematic ranges of the dataset, tested based on comparison can be used. In this test, the adoption of different sampling strategies affects the time and cost of the tests. Once both the reference data and test data are available, it is necessary to determine the data comparison method. For example, the test of positional accuracy is recommended comparing the coordinates of points within the testing dataset to the coordinates of the same points from reference dataset. The operation of data-matching can be performed either automatically or interactively, which aims to determine the differences between the testing data and reference data.

Test by deductive estimate is based on knowledge of errors in data production process. This method determines the quality results using rules of error propagation. For example, the accuracy of a vector-based dataset can be estimated by accumulating the errors of data conversion, data manipulation and data process.

Test by inference assumes a map series that subjects to same production methods has the same accuracy level, so by knowing the accuracy of one map, the accuracy of all subsequent maps in the same series can be inferred.

Specific logical and topological check is a kind of automatic computer program check. This method tests the logical consistency of dataset, for example, all chain intersect at nodes, self-interested polygons, and inner rings embed consistently in enclosing polygon, etc. Moreover, the topological relationship between geographic data can be verified by a set of rules defined by operator, for example, river should not be situated on the surface of road.

2.4.1 Test for Positional Accuracy

In vector-based spatial data, positional accuracy includes horizontal accuracy and vertical accuracy of point, line and polygon features. Error models of point feature have been well developed for the applications of surveying and Geographic Information System (GIS). Positional accuracy of point is presented by an error ellipse or error rectangle. Positional error models of line have been developed by researchers such as Dutton G. (1992), Shi (1994, 1998, 2000) and Goodchild and Hunter (1997), etc. According to these models, positional accuracy of line can be estimated using law of error propagation, statistical approaches and analytical approaches based on the errors at two end points.

Ariza-Lopez F.J. and Mozas Calvache A.T (2012) conducted a study comparing four line-based positional assessment methods, which are: Hausdorff distance method, mean distance method, single buffer overlay method and double buffer overlay method. The result showed that the single buffer overlay method proposed by Goodchild and Hunter (1997) gave a more general solution because it included the other's results.

This study reviews three dominant error models for line feature, which are single buffer overlay, confidence region models and a stochastic process-based model for the positional error of line segments.

2.4.1.1 Single Buffer Overlay

Goodchild and Hunter (1997) introduced a method to assess the accuracy of line segment by measuring the percentage of the length of tested line within buffers drawn around a reference line. Figure 2.8 shows an example of tested line and reference line for the test. The buffers are used to trace the line segments that are contained in a specified buffer width. Assume a buffer width is 0.5 m around the reference line, the length of tested line within the buffer polygon can be determined. Then, this length is compared to the total length of tested lines to compute the proportion *(p)* of line segments within 0.5m buffer distance. This method is relatively insensitive to outliers and rare situations.

Let *p(x)* is the function of the proportion and *x* is the buffer width.

$$p(x) = \frac{\text{length of tested line within a buffer}}{\text{total length of tested line}}$$

Assume *p(x)* is a cumulative probability distribution with *p(x)* = 0 and $p(\infty)$ =1, then values of x can be thought of as corresponding to percentiles of the distribution. For example, the 95th percentile as the distance within which 95% of the length of the tested line lies within the buffer polygon.

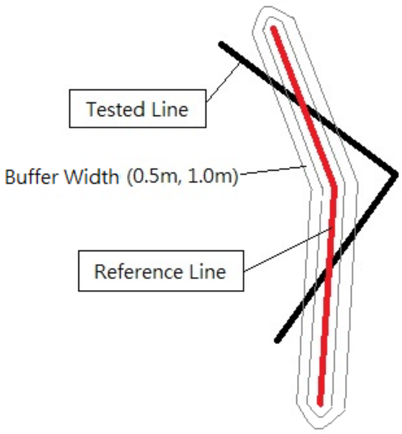


Figure 2.8 A buffer width (0.5m, 1.0 m) around the reference line is intersected with the tested line, to determine the percentage of tested line lying within the buffer distance

The buffer width can increase until the corresponding buffer width with respect to the target percentage (i.e. 50%, 60%, 70%, 80%, 90% and 95%) is found. Furthermore, Goodchild and Hunter (1997) proposed a procedure to find the target percentile which is a simple way to obtain positional accuracy for linear feature in spatial dataset and is easy to implement in GIS.

2.4.1.2 Confidence Region Models for Line Features

This model defines a line segment in two-dimensional space as a region that contains the true location of the line segment with a predefined confidence level. Figure 2.9 shows the confidence region J_2 for a measured location $Q_{21}Q_{22}$ of the line segment with a predefined confidence level $\alpha = 0.97$. Points Q_{21} and Q_{22} are measured locations of the endpoints of the line segment, while points $\emptyset_{21}$ and $\emptyset_{22}$ are their corresponding true locations. This model assumes that error in the two end points is identical and follows the normal distribution. The confidence region of line segment is affected by the positional error at the end points and the predefined confidence level.

When the positional error at end points is larger, the line segment confidence region is wider.

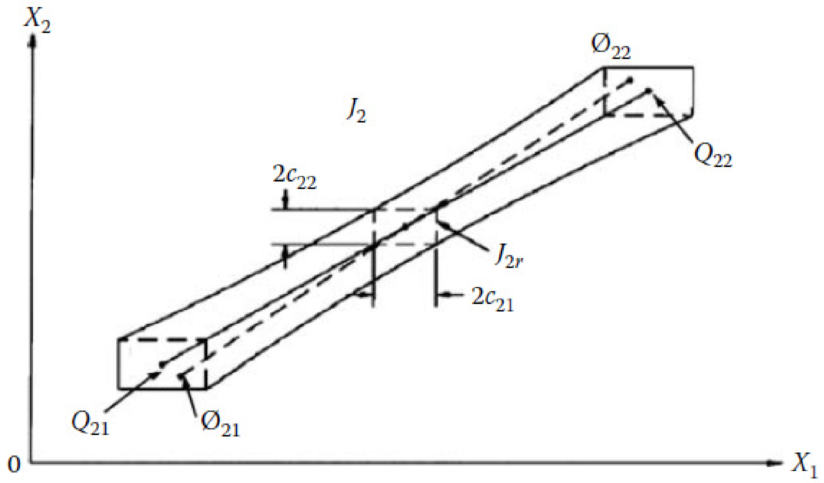


Figure 2.9 The confidence regions of a line segment in two-dimensional space (Shi W., 1998)

2.4.1.3 A Stochastic Process-based Model for the Positional Error of Line Segments in GIS

Shi W. and Liu W. (2000) developed an error-band model (G-band model) describing the positional error of line segments. This model is based on the theory of stochastic process and has two assumptions: a) two endpoints of a line segment are auto-correlated and cross-correlated, and b) two endpoints follow two-dimensional normal distribution. The uncertainty information matrix of line segments is derived to indicate the error of an arbitrary point on the line segment. The distribution and errors of an arbitrary point on the line segment can be given in this error model.

Line Segment Description:

Figure 2.10 shows a line segments Z_0Z_1 which composites of two endpoints Z_0 and Z_1. $Z_t(X_{(t)}, Y_{(t)})$ is an arbitrary point on the line segment Z_0Z_1. The two end points are represented by the vectors $Z_0(X_0, Y_o)^T$ and $Z_1(X_1, Y_1)^T$. This model is assumed generic. In addition, the error of a point in spatial data may be caused by many error sources, such as uncertainty inhered in the original paper map, uncertainty from digitization and uncertainty in data processing or editing, so it can assume that:

$$Z_i \sim N_2(\mu_{Z_i}, \Sigma_{Z_i Z_i}) \quad (i \in [0,1]) \qquad (2.4.1.3.1)$$

where:

$\mu_{Z_i} = (\mu_{x_i}, \mu_{y_i})$ is the mean value vector of the arbitrary point on the line segment

$\Sigma_{Z_i Z_i}$ is the variance-co-variance matrix of the point. It can be represented as:

$$\Sigma_{Z_i Z_i} = \begin{bmatrix} \sigma_{x_i}^2 & \sigma_{x_i y_i} \\ \sigma_{y_i x_i} & \sigma_{y_i}^2 \end{bmatrix} \qquad (2.4.1.3.2)$$

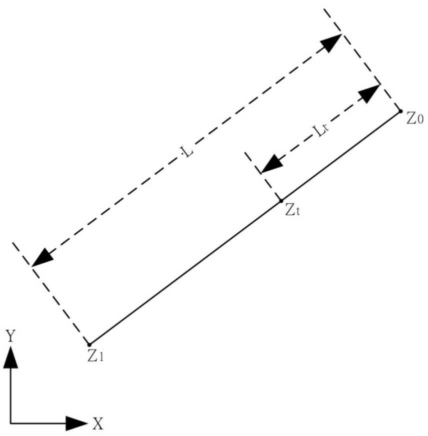


Figure 2.10 The line segment Z_0Z_1 composites of the two endpoints $Z_0(X_0, Y_0)$ and $Z_1(X_1, Y_1)$

(Shi W. and Liu W., 2000)

The coordinates of the arbitrary point on the line segment can be represented as:

$$\begin{cases} X(t) = (1-t)X_0 + tX_1 \\ Y(t) = (1-t)Y_0 + tY_1 \end{cases} \quad (0 \leq t \leq 1) \qquad (2.4.1.3.3)$$

The Error-band Model:

In the G-band model, the shape and size are dependent on the shape of the spatial density distribution function surface. According to (2.4.1.3.3), an arbitrary point $Z_t(X(t), Y(t))$ on the line segment follows two-dimensional normal distribution. The major semi axis, minor semi axis and the direction to the major semi axis of the corresponding error ellipse of an arbitrary point on the line segment can be derived as:

$$\begin{cases} A^2(t) = \sigma_0^2 \lambda(t)_1 = \frac{1}{2}\{\sigma_X^2(t) + \sigma_Y^2(t) + \omega\} \\ B^2(t) = \sigma_0^2 \lambda(t)_2 = \frac{1}{2}\{\sigma_X^2(t) + \sigma_Y^2(t) - \omega\} \\ \tan 2\varphi(t) = \frac{2\sigma_{XY}(t)}{\sigma_X^2(t) - \sigma_Y^2(t)} \end{cases} \qquad (2.4.1.3.4)$$

$$\text{where } \omega = \sqrt{\sigma_X^2(t) - \sigma_Y^2(t) + 4\sigma_{XY}^2(t)}$$

Referring to (2.4.1.3.4), an infinite number of error ellipses of an arbitrary point can be drawn on the line segment Z_0Z_1. Based on the stochastic process, the nature of positional error of a line segment can be completely described by using this class of error ellipses. The boundary lines and error ellipses of the end points construct the band-shape area around the true or mean value of the line segment. This is named the G-band model.

Figure 2.11 shows the visualization of G-band of a line segments with the geometric parameters calculated from formula (2.4.1.3.4). The ratio between semi-major radius and semi-minor radius is 2:1. Examples (1) and (2) of figure 2.11 are directional independent error ellipses and (1) is actually an error-band. Example (4) shows homogenous error ellipses, and examples (5) to (8) show sets of error ellipses of general line segments. This G-band model provides the analytical relationships between the band shape and size with the errors of the two end points and relationships between.

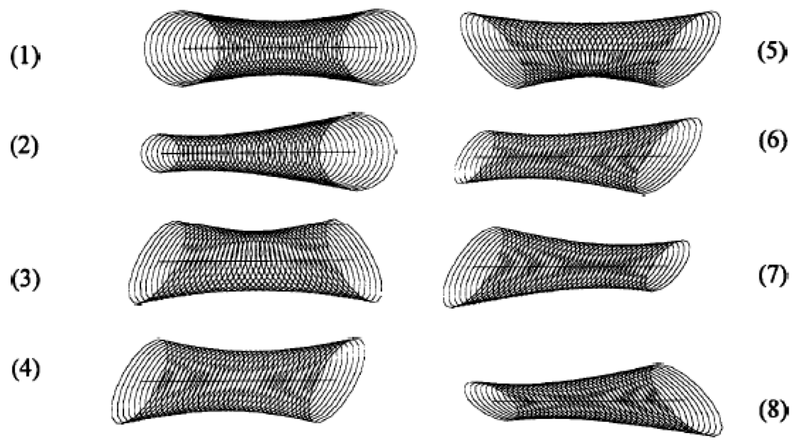


Figure 2.11 Visualization of G-band of the random line segment Z_0Z_1

(Shi W., Liu W., 2000)

2.4.2 Test for Completeness

Completeness error can be classified into commission error and omission error. In most GIS applications, completeness error can be applied to the level of graphical display, the data, the data relationship and the attributes. Field survey, graphical check and automatic program check are common methods to detect completeness error. By the field surveying, "true" value (or reference value) of testing feature such as positional information and thematic information can be obtained, and considered as reference data to compare with those in topographic data to find out whether there are excess or less. Inspection of graphical display, either on survey sheet or computer display, checks the completeness in term of graphical representations. For example, the adequacy on spot level over a wide area, especially around the abrupt change in gradient, can be easily identified. Automatic program check mainly includes database inspections. Through programs batch checking, missing or redundancy of attribute data, relationship and domain can be controlled.

2.4.3 Test for Thematic Accuracy

Thematic accuracy includes classification accuracy, quantitative attribute accuracy and qualitative attribute accuracy, for example, errors in street name, feature code, building type and building name, etc. Inspection of thematic accuracy is usually performed by field survey and automatic program check. In field survey, the attribute values in spatial data are compared with the corresponding values in the field. In automatic program check, attribute filtering can be used to select errors in database.

2.4.4 Test for Temporal Accuracy

Temporal accuracy is an important factor for topographic data because it provides users information for judging whether the spatial data is updated or not. Temporal accuracy concerns the dates of data acquisition, types of updates and validity periods. In most dataset, the temporal information has been treated as an attribute, such as revision date or data-input date, separately stored in the dataset, and sometimes modeled as a date, an interval, or a temporal range. Test of temporal accuracy requires historical records, for example, aerial photograph, old survey sheet and archive data. To maintain the temporal accuracy of dataset, specific mechanisms should be established to allow version-management in data creation, data modification and data deletion.

2.4.5 Test for Logical Consistency

Logical consistency relates to all logical rules that govern the conceptual, domain, topologic and format of geographic data. Automatic program check is the most efficient method to evaluate the logical consistency of both spatial and non-spatial data. A topographic dataset called logical consistency if it respects to the structural characteristics of data schema and database design. For example, all chains intersect at nodes, topology relationships are represented and respected, the data file is consistent, and the variables used adhere to the appropriate value (string, double, type, etc.).

2.5 Provision of Data Quality Information of Topographic Data

Vector-based topographic data has been widely used and shared in private and government agencies to support asset management, planning, and environmental analysis. The advantages of easily customization and modification make vector-based topographic data the common data format in GIS applications. With enormous potential users, the provision of precise and concise data quality information, and perhaps satisfying the requirements of most uses, is important (Harding J., 2006).

This section analyzes the data quality information provided by three national mapping agencies, which are: a) U.S. Geological Survey, b) Geoscience Australia, and c) Ordnance Survey. The analysis is summarized in table 2.8. The quality information presented in the table is extracted from following three product standards, specifications or user guide.

a) Digital Line Graph Standards – Department of the Interior U.S. Geological Survey National Mapping Division (DLGS, 1999)

b) Geoscience Australia Topographic Data and Map Specifications for the National Topographic Database & NTMS Series 1:250,000 & 1:100,000 scale Topographic Map Products –Geoscience Australia (NTMS, 2012)
c) OS MasterMap Topography Layer User Guide – Ordnance Survey (OSMM, 2016)

Table 2.8 Summary of quality information for three topographic datasets

	Digital Line Graphs (DLG) (U.S. Geological Survey)	**The National Topographic Database (NTDB) (Geoscience Australia)**	**OS MasterMap Topography Layer (OSTL) (Ordnance Survey)**
Product Scale	Large scale: - 1:20,000 - 1:24,000 - 1:25,000 Intermediate scale: - 1:100,000 Small scale: - 1:200,000 Other scale: - 1:48,000 - 1:50,000 - 1:62,500 - 1:63,360	Database scale: - 1:25,000	Urban area: - 1:1,250 Rural area: - 1:2,500 Mountain / Moorland area: - 1:10,000
Revision Policy /	- Adopt revision cycle policy	- adopt a “change only” basis revision policy	- adopt continuous revision and cyclic

Production Cycle	- Each revision cycle is about 3 years	- Those changed features (add, update and delete) will constitute the information supplied	revision - Continuous revision is conducted within six months of construction being completed - Cyclic revision is conducted periodically mainly for changes to the natural environment - All designated prestige sites are captured before they are open to public
Lineage	Source of the data file: - Of the 1:24,000 scale data collected from 15 minute quadrangles, the majority are digitized as four 7.5 minute units, and distributed in standard 7.5 minute cells Reference system: - UTM - Albers Conical Equal Area -	Data projection: - GDA94 UTM Precision: - Double precision XY Resolution: - XY resolution is “0.0000005” degrees XY Tolerance: - XY tolerance is “0.0000001” degrees	Coordinate reference system: - UTM Resolution of seamless orthorectified aerial images: - 25 cm
Positional	- Source graphics are	- Positional accuracy is	- For Scale 1:1,250

<table>
<tr>
<td>Accuracy</td>
<td>normally complied to meet NMAS (1947), where 90% of well-defined feature are to be within 0.02 inches of true mapped ground position
- The DLG positional error shall be less than or equal to 0.003 inches standard error in both x and y component directions, relative to the source that was digitized
- Edge alignment in node snapping are within 0.02 inches, and maximum positional adjustment for any node would be 0.01 inches</td>
<td>estimated by modelling the propagation of errors in the data production process or by directly comparing the coordinate locations in the completed data against a source of significantly higher known accuracy.
- The absolute planimetric accuracy is “Not more than 10% of points will be in error by more than 140m for 1:250,000 data, 56 m for 1:100,000 data and 14 m for 1:25,000 data
- The absolute vertical accuracy is ±5m in spot elevation
- The accuracy of the contours is defined as 1/2 of the contour interval, for example, ±10m for a 20 m contour interval</td>
<td>
<table>
<tr><td></td><td>At 95% Confidence Level</td></tr>
<tr><td>Absolute Accuracy</td><td>0.8 m</td></tr>
<tr><td>Relative Accuracy</td><td>±0.9 m (up to 60m)</td></tr>
</table>
- For Scale 1:2,500
<table>
<tr><td></td><td>At 95% Confidence Level</td></tr>
<tr><td>Absolute Accuracy</td><td>1.9 m</td></tr>
<tr><td>Relative Accuracy</td><td>±1.9 m (up to 100m)</td></tr>
</table>
- For Scale 1:10,000
<table>
<tr><td></td><td>At 95% Confidence Level</td></tr>
<tr><td>Absolute Accuracy</td><td>7.1 m</td></tr>
<tr><td>Relative Accuracy</td><td>±7.7 m (up to 500m)</td></tr>
</table>
</td>
</tr>
</table>

Attribute Accuracy / Thematic Accuracy	- All attribute codes will agree within 98.5% to attribute codes described in the standard	- Items will be populated in accordance with the population requirement codes set out in Data Attributes Rules for each feature type	- N/A
Edge Matching	- Edge matching ensures that features are matched in content, position and attribution along a common edge	- Linear and polygon features should be spatially join across the limit of data feature. The distance for edge matching will be 140 m for 1:250,000 or smaller scale, 56 m for 1:100,000 and 14 m for 25,000	- N/A
Edge Align Status and Reason Flags	- Information in the header indicates the status of the file with respect to the edge matching result	- N/A	- N/A
Logical Consistency	- Geometry checks have been applied to dataset, such as lines begin and end at node, and lines connect to each other at nodes, etc	- Geometry checks have been applied to dataset, such as no overshoots and no undershoots, etc	- N/A
Completeness	- For a given category of data will contain at least	- N/A	- N/A

	the same level of content and detail shown on the source graphic		
Temporal Accuracy	- N/A	- N/A	- N/A

The above comparison shows that revision policy, lineage, positional accuracy, attribute accuracy and logical consistency are the most important quality information that should be incorporated into the product specifications, user guides or content standards. This implies that they are the major factors for users to decide whether a topographic dataset does fit their intend use.

As there are many GIS software provides automatic functions to detect logical consistency errors such as edge matching, overshoot and undershoot, etc., data providers tend to conduct logical consistency check during data editing and data processing. This ensures the final product is free of logical consistency error.

DLG and NTDB still adopt the principles of NMAS (1947) to describe the positional accuracy, while OSTL adopts the NSSDA (1998) with 99% and 95% confidence level. In the digital era, most digital topographic data use the concept of resolution to refer to the precision of data, not the scale. Therefore, reporting the positional accuracy using NSSDA (1998), i.e. 95% confidence level of both horizontal and vertical accuracy, is more appropriate and well known by most users.

2.6 Quality Control of Topographic Data

To maintain the quality of topographic data, implementation of quality controlling system throughout the entire data production process is the key. Standardization also plays a dominant role, because most topographic data are created according to widely accepted standards, such as positional accuracy, logical consistency and completeness, etc. This section describes the principles of quality control in data production process and reviews the basic components of spatial data standards.

2.6.1 Quality Control in Lifecycle of Topographic Data

The production of topographic data is a process from identification of user's need to data dissemination. This includes five major steps: a) Identify user's requirements, b) specify product specifications, c) data production, d) QA/QC of product, and e) data dissemination. At each steps, it is required to check whether the results of previous steps are satisfactory. These steps are organized as a continuous cycle for production of topographic data and form a quality control framework. (Dassonvile L., Vauglin F, Jakobsson A. and Luzet C., 2002). Figure 2.12 presents the lifecycle of topographic data.

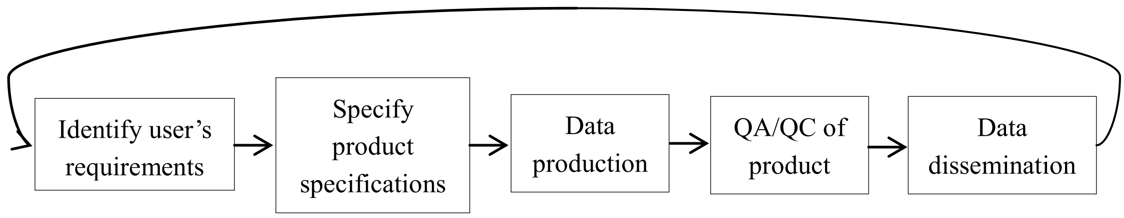


Figure 2.12 Lifecycle of topographic data

2.6.1.1 Identify User's Requirements

Understanding the needs of user is the key to uphold a high quality topographic dataset. There are various ways to obtain user's needs and feedbacks, such as market research questionnaires, on-line enquires services, consultations with users, setup a customer services section and conducting an analytical research on the user's requirement, etc. This information is critical for data providers to plan the future developments on data usage, data capture, database design and production specifications.

2.6.1.2 Specify Product Specifications

Product specifications define the abstraction of real world objects and provide definitions for both geographical features and attribute information stored in database. It gives real world objects a conceptual model that is suitable for geographical representation and analysis. Some product specifications include the data capture standards, such as positional accuracy requirements, data conversion methods, data capture methods, and data quality information, etc. Moreover, product specifications are reference document for internal quality assessment, because the conformance level of a dataset is usually contained in the specifications, such as horizontal accuracy and vertical accuracy.

2.6.1.3 Data Production

Quality control in data production requires sufficient knowledges in different data capture methods. Section 1.1.1 has introduced the error sources of common data capture methods. A comprehensive data acquisition plan, which includes calibration records, accuracy of control network, data capture criteria and site circumstance, is significance in controlling the quality of data production. During data input and data processing stage, customized program and software can assist to detect human errors, such as typo, overshoot, undershoot and edge matching, etc., so it is suggested incorporating these functions into the data editing stage.

2.6.1.4 QA/QC of Product

The quality assessment methods of vector-based topographic data have been summarized in section 2.4. These methods can be used in data creation and data updating to ensure the quality of data fulfills the requirements stated in product specification. During the QC/QA process, the acceptance or rejection criteria should be defined in advance according to the potential applications of data.

2.6.1.5 Data Dissemination

Data dissemination is the final steps of lifecycle of topographic dataset. The final product together with data quality information is delivered to users. If users give comments on the product, it will go to first steps.

2.6.2 Data Standards

Data needs to be captured and processed carefully, and organized according to defined rules and product standards. The establishment and publication of data standards plays an important role in data quality control, and is essential for users and data producers to have an overview in the content of product. The Content Standard for Digital Geospatial Metadata of the Federal Geographic Data Committee (CDGM, 1998) is an example that has been widespread accepted and used in the development of data standard.

Fitness for use and easily understanding are key elements of data standards. Good data standards are developed in an inclusive way and are widely used. Data standards contain three main components: a) standard for information sharing, b) spatial standard, and c) non-spatial standard.

2.6.2.1 Standards for Information Sharing

Communication of data standards should be employed to permit the information to transmit freely through internet or other networks. The communication standards are recommended to be open standards that are widely accepted and implemented by Government and private sector. The communication protocols, such as Hypertext Transfer Protocol (HTTP), Extensible Markup Language (XML), Geography Markup Language (GML), and Simple Object Access Protocol (SOAP) are examples of such standards.

2.6.2.2 Spatial Standards

Spatial standards are used to describe the location related information for spatial data. This includes standards associated with scales, accuracy, datum, map projections, coordinate system, etc. This information is essential for describing the spatial features. For example, the topographic data cannot be used unless it contains coordinates transformation parameters and projection information.

2.6.2.3 Non-spatial Standards

Non-spatial standards include information related to data captures, process, storage, maintenance, presentation, dissemination and quality level, etc. For example, the data quality information is represented in terms of completeness, positional accuracy, logical consistency, temporal accuracy and attribute accuracy, etc.

CHAPTER 3

PROPOSED PROCEDURE FOR DATA QUALITY ASSESSMENT

Vector-based topographic dataset may contain data that obtained from different sources, single quality evaluation method and sampling method may not fulfill all requirements in quality assessment. To assess the quality of vector-based topographic data, an efficient and effective procedure, including the quality assessment methods, sampling methods and reporting methods, must be adopted. Figure 3.1 shows the proposed quality assessment procedure. The entire data quality assessment procedure contains four steps, and the details of each step are illustrated in this section.

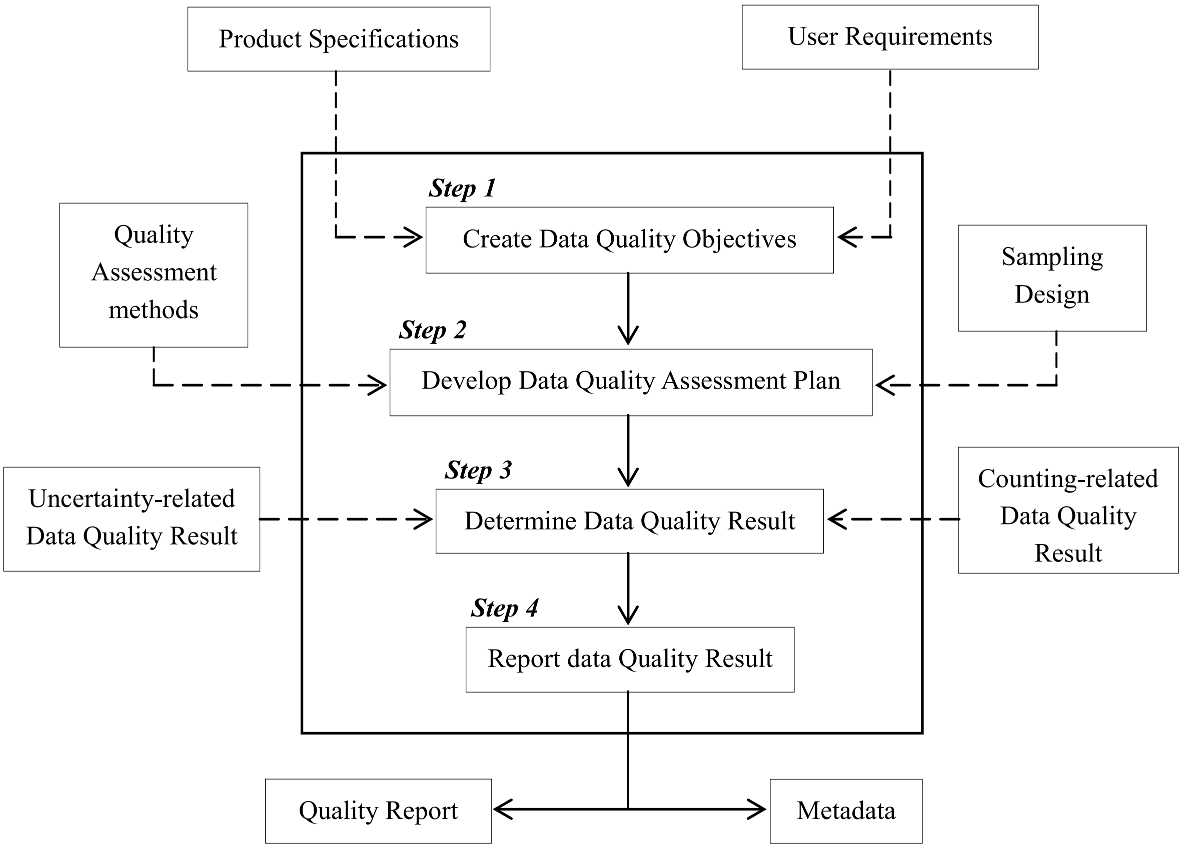


Figure 3.1 Proposed quality assessment procedure

3.1 Step 1: Create Data Quality Objectives

Data quality is classified into internal quality and external quality. Internal data quality is the degree to which the data conforms to the product specifications. External data quality is the degree of how much the data fulfills the needs of users. Both internal quality and external quality are described using data quality elements and data quality subelements, for example, completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy. As new type of spatial data, such as transportation network, address dataset and 3D model, has been used in many GIS applications, there is a need to create additional data quality elements or subelements to describe their quality. An approach to create data quality objectives is presented in this section and it includes two components: a) identify product requirements, and b) creation of new data quality elements.

The creation of data quality objectives is initial stage of the entire data quality assessment procedure. It gives an overview to the quality assessment procedure. The objectives are mainly derived from requirements stated in product specifications, reported errors and comments on the dataset. Thus, the identification of internal and external product requirements is ideal methods to obtain the quality objectives. Three examples of data quality objectives are presented in table 3.1.

Table 3.1 Examples of data quality objectives

Example no.	Data Quality Objectives	Data Quality Elements	Data Quality Subelements	Descriptions of Data Quality Objectives
1	The absolute horizontal accuracy of building features	Positional accuracy	Absolute accuracy	Compute the RMSE from the error distances
2	All buildings classified as “Police Station” shall bear address record(s)	Completeness	Omission	If a building classified as “Police Station” and does not bear building address, it is an error
3	For adjoining 3D RoadPolygons, the vertices with same x, y coordinates should have same z coordinate	Logical consistency	Connectivity error *(new quality subelements)*	If two RoadPolygons are adjoined and the common vertices that have same x, y coordinates but have different z coordinates, the two RoadPolygons are considered having connectivity error

3.1.1 Identify Product Requirements

Based on the principles of spatial data quality, the product requirements are divided into internal requirements and external requirements. The internal requirements are derived from the product specifications, while the external requirements are acquired from the user's feedback on the data.

3.1.1.1 Internal Quality Requirements

Data providers tend to produce several series of topographic map for various applications. For instance, the Survey and Mapping Office (SMO) of Hong Kong Special Administrative Region (HKSAR) has produced four series of digital topographic map with different scales. Each map series has individual product specification, data model, and user guide. To identify the internal requirements, a map series is divided into different level of details in order to benefit to the quality assessment and reporting. Figure 3.2 shows the data structures of topographic data and different level of details.

For the purpose of identifying the internal quality requirements, clearly understanding the major usages of the dataset is significant. This enables data producers to establish the extent to determine internal quality requirement. In addition, a comprehensive product specifications and data standards, including clear statements on acceptance criteria and threshold valve, can be used to define the error level and conformance level.

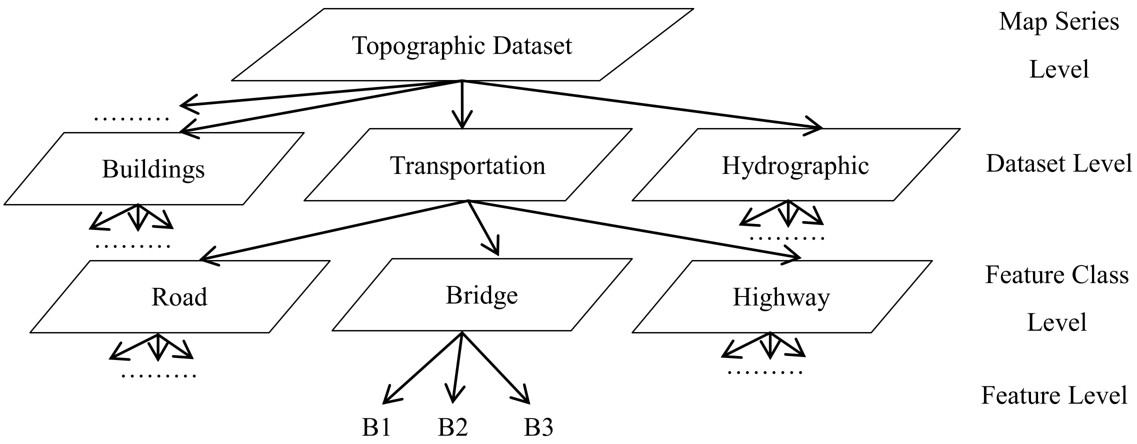


Figure 3.2 Data structures topographic data

For the map series level, the data quality requirements can be obtained from the general purposes and usages of dataset. For example, the positional accuracy requirement can be inferred from the map scale. If the map scale is 1:1,000, then a plotting accuracy of 0.5 mm would be represented 0.5 m on the ground. This is the general positional accuracy requirement of the map series. The map series level requirements are the highest level of the dataset and usually considered as the major quality information representing in product user guides and content standards.

The dataset level is a grouping of features classes that with similar natures or characteristics. For instance, the OS MasterMap Topographic Layer contained nine data themes, which are administrative boundaries, buildings, heritage and antiquities, water, land, rail, road, track and paths, structures, and terrain and height. The individual data theme would have quality requirements that suit for specific purpose. For example, the road theme would have higher quality requirements in completeness and temporal accuracy than other theme, because it needs to support the road network applications.

The feature class level is a grouping of features that represent the same type of objects in the real world. Building is a typical example of feature class which exists in most national map series. This level may contain detailed descriptions and data capture requirements for the real world features. For example, SMO defines building as a rigid, fixed and permanent structure that is roofed and walled for the support, shelter or enclosure of people, animals or property. The building feature with the size larger than 4 m^2 should be captured. This information constitutes a set for internal requirements for data quality assessment.

The feature level is a single feature that describes an individual objects in the real world. In vector-based topographic data, features can be represented by point, line and polygon. Some of these features may have specific usage, for example, building features with feature type equal to "Police Station" must contain address information. This specific requirement for police station forms the internal requirements of dataset.

3.1.1.2 External Quality Requirements

External quality requirements relates to the concept of "fitness for use", which is the level of acceptance between the data and the user's expectations. Once the data has been delivered to the market, data providers can evaluate the "fitness of use" based on the experience of users or the comments received from experts in the field of applications. Figure 3.3 presents an approach to identify external quality requirements.

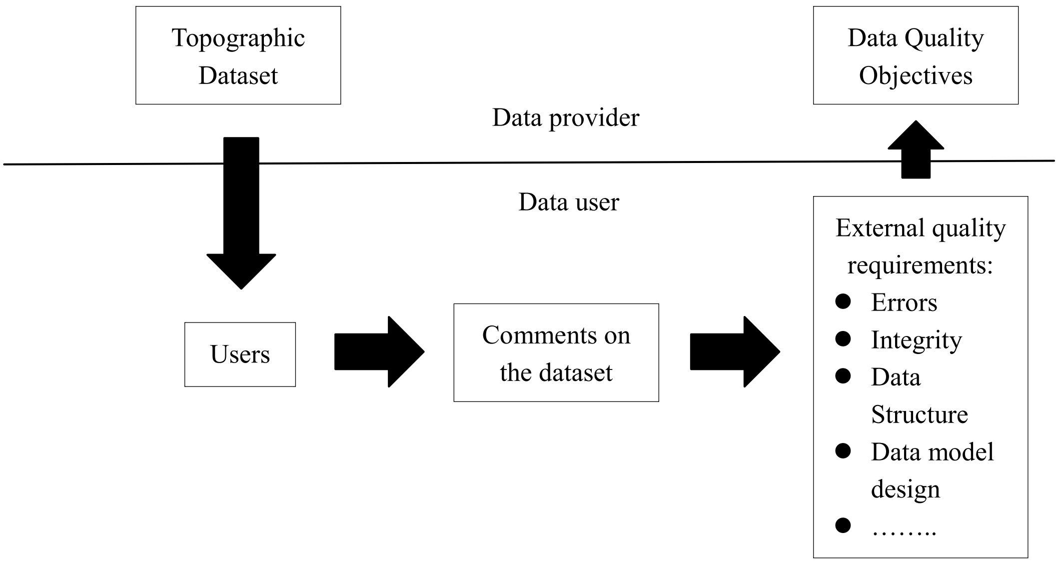
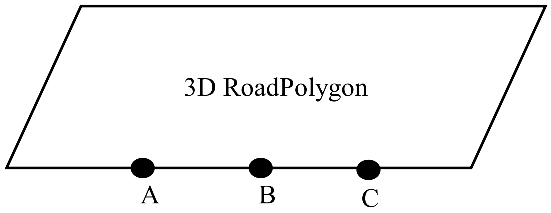


Figure 3.3 An approach to identify external quality requirements

After received the comments on the dataset, data provider can consolidate the information and categorize the conceptual problems into different aspects for data quality assessment, such as completeness, attribute accuracy, logical consistency and temporal accuracy, etc. For instance, a data user reported that the utility number stored in the database is inconsistency with the real world objects. Data provider can classify this problem as a matter of attribute consistency and conducts a quality assessment on utility numbers for estimating the error level for whole dataset.

3.1.2 Creation of New Data Quality Elements / Subelements

If the data quality elements and data quality subelements defined in ISO 19113 or other national standards are inadequate to describe the quality of a product, addition data quality elements or data quality subelements shall create for quality assessment. This may happen when new products arise together with its quality information has to be reported. In this circumstance, the definitions of new data quality elements or subelements should be clearly defined in data quality report in advance. Figure 3.4 shows an example of new data quality subelement which describes the quality of 3D RoadPolygon.



Point A, B and C are vertices of 3D RoadPolygon. Line AB and Line BC are two continuous line segments, the gradient change of Line AB and Line BC cannot exceed a threshold “0.15”.

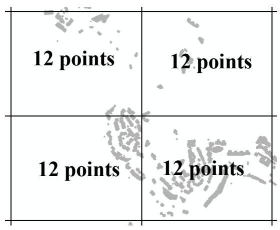
Figure 3.4 An example of new data quality subelement

3.2 Step 2: Establish Data Quality Assessment Plan

The quality assessment plan proposed in this section incorporates two parts: a) data quality assessment methods and b) sampling plan. Section 2.4 has reviewed several data quality assessment methods, and the results showed that most national standards prefer using direct assessment method, including external and internal assessments, to conduct quality assessment. For external assessment, the methods for positional accuracy assessment and attribute accuracy assessment are introduced. Among these methods, the quality of reference data or the definitions of “well-defined” features will directly influence the results. This section illustrates the principles in handling these issues. For internal assessment, customized software and commercial off-the-shelf (COTS) functions are adopted to detect the errors. This section analyzes their applications and flexibilities.

Sampling process is the means to accomplish the direct quality assessment, and it requires testing sufficient features or attributes in the test data to achieve the data quality result. Based on the data quality objectives, a sampling plan is proposed. Table 3.2 shows three examples of data quality assessment plans which using the data quality objectives defined in table 3.1.

Table 3.2 Examples of data quality assessment plans

Example no.	**Features or Attributes to be Tested (Refer Table 3.1 - Data Quality Objectives)**	**Quality Assessment Methods**	**Sampling Methods**	**Sampling Size or Sampling Ratio**	**Spatial Distribution of the Samples**
1	Building features	For each building feature, measure the error distance between absolute coordinate values ("well-defined" building corners only) in test data and those in reference data, and compute the RMSE at 95 % confidence level for the error distance. The reference data shall be collected by field	Area-guided sampling	50	 Remarks - the outermost rectangle is the total area to be assessed (population), and it is divided into four small rectangles with equal area. The number of sampling points within each

		survey with absolute horizontal accuracy better than 0.3 m (three times of the test data) at 95% confidence level			small rectangle is "12".
2	All Buildings classified as "Police Station"	Select all buildings that classified as "Police Station" and check whether they bear address record(s). An automatic program check is utilized to assess the number of errors in entire database.	Full Inspection	N/A	N/A
3	All adjoining RoadPolygons	Select all adjoining RoadPolygons using spatial query. An automatic program check is used to assess the percentage of errors in entire database	Full Inspection	N/A	N/A

3.2.1 Data Quality Assessment Methods

Two data quality assessment methods are presented: a) external assessment, and b) internal assessment. External assessment compares the difference between the reference data and test data, while internal assessment evaluates the data internally, for example, domain consistency and edge matching, etc.

3.2.1.1 External Assessment

Vector-based topographic data contains two major elements: a) geometries, which are point, line and polygon, and b) attributes, such as textual information, building name and address, etc. To assess the quality of geometry features (i.e. positional accuracy), comparison of point or line features within the dataset to the same features from an independent source of higher accuracy is the preferred method. In following section, we term the first representation the "test data" and the second representation the "reference data".

To assess the quality of attributes (i.e. attribute accuracy) the information, such as feature codes, textual information and quantitative value, in test data are compared with the same attribute information in reference data. The methods of attribute accuracy assessment are demonstrated.

3.2.1.1.1 Positional Accuracy Assessment

The methods for positional accuracy assessment have been reviewed and analyzed in section 2.2 and 2.4.1. These methods proposed that the positional accuracy is tested by comparing the coordinates of well-defined points in the test data to the coordinates of the same points from reference data. Well-defined features are those that easily visible or recoverable in reference data and test data, such as building corners, intersections of roads and lamp posts, etc. The reference data shall be acquired separately from the test data, and supposes to be three times more accurate than the expected accuracy of the test data.

Although the methods for positional accuracy assessment have been well addressed in the above literature review, there is still a room for further discussion. The method proposed in this part aims to deal with two problems: a) how to select well-defined features in test data and reference data, and b) how to deal with the problem when the accuracy of reference data is the same as the test data.

In selecting well-defined features in test data, the effect of generalization shall be taken into account. Figure 3.5 shows an example of generalization of building feature in a 1:1,000 topographic dataset. The black polygon represents a building feature. According to the product specifications, the jut or recess of building less than 1.0 m is ignored in cartographic representation and the longer portion of the building is adopted. Since the horizontal accuracy standard for this dataset is 0.3 m, the selection of "correct" building corners would be the major factor affecting the result in positional accuracy assessment. In case there are no well-defined points identified in the polygon features, the method of single buffer overlap mentioned in section 2.4.1.1 can be used. Figure 3.6 (a) and (b) shows two examples of well-defined points in line features and polygon features. They are often shape building corners without generalization, or shape nodes or vertices in road margin.

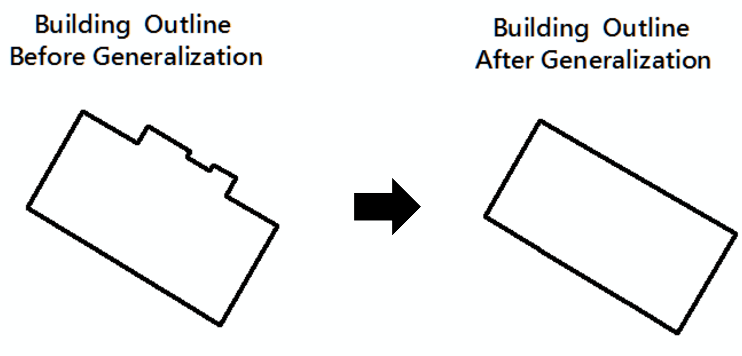


Figure 3.5 An example of generalization for cartographic presentation

(a) (b)

Figure 3.6 (a) The polygons is building feature, all building corners can be identified in orthophotos, (b) The lines are road margin, there are well-defined vertices identified on orthophotos

Considering the accuracy of reference data that may not fulfill the testing requirement (i.e. three times accurate than the test data), the error in reference data shall be added into the result during positional accuracy assessment. In these cases, RMSE at 95 % confidence level is used to determine the accuracy of point features. Confidence region model (Shi, 1994) is used to determine the accuracy of line features.

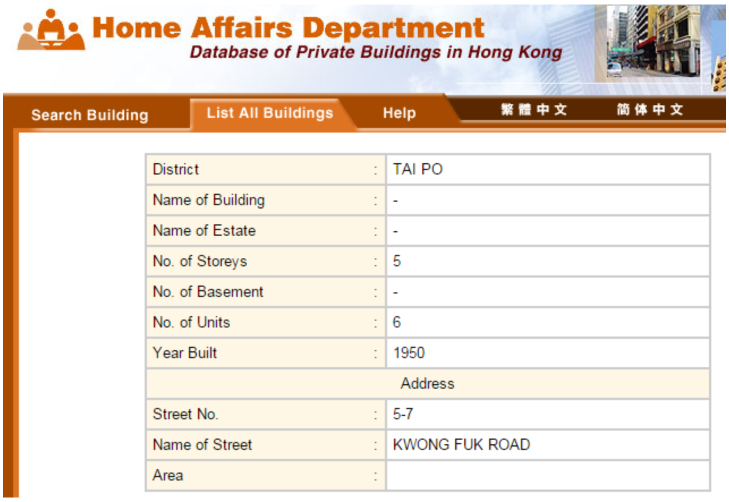
3.2.1.1.2 Attribute Accuracy Assessment

The attributes defined in this section are building names, addresses and textual information. These attributes often contains errors, because:

a) Some of them are inputted by operators. Careless mistakes are difficult to control.
b) The building name has been changed in real world, but the database cannot be updated accordingly.
c) Textual information always stored using UNICODE. For Chinese characters, it is easy to misidentify some Chinese characters which are similar.

d) Some attributes are stored in a separate table in database, the relationship between the geometry feature and table records cannot be maintained.

To assess the quality of attribute information, the official or government records can be adopted as reference data for checking. Figure 3.7 shows a database of private buildings in Hong Kong. The building names and address records stored in test data can be compared to the same records in official dataset for attribute accuracy assessment.


Home Affairs Department
Database of Private Buildings in Hong Kong
Search Building | List All Buildings | Help | 繁體中文 | 简体中文

District	TAI PO
Name of Building	-
Name of Estate	-
No. of Storeys	5
No. of Basement	-
No. of Units	6
Year Built	1950
Address	
Street No.	5-7
Name of Street	KWONG FUK ROAD
Area	

Figure 3.7 Database of private buildings in Hong Kong

Available at: https://bmis1.buildingmgt.gov.hk/bd_hadbiex/home.jsf?lang=tc

[Accessed 3 May, 2016]

3.2.1.2 Internal Assessment

File geodatabase is a common format that adopted by national mapping agencies, such as Geoscience Australia, Ordnance Survey and SMO in Hong Kong, to store vector-based topographic data. In traditional database, tile-based structures incurred errors such as edge matching. The file geodatabase stored geographic features in seamless structure, thus the problems of edge matching are relieved. However, there are still some internal problems. Table 3.3 summarizes the possible internal errors that will be appeared in file geodatabase. The errors mentioned in table 3.3 can be detected by commercial off-the-shell functions or customized program automatically.

Table 3.3 Summary of internal error in file geodatabase

Type of Problems	**Possible Internal Errors**
Database	- Domain inconsistency - Invalid subtypes - Invalid relationships
Geometry	- Invalid geometry - Multi-part line - Multi-part polygon - Nonlinear segment - Duplicated geometry - Duplicated vertices
Polygon	- Silver polygon - Invalid hole feature - Polygon area are too small
Polyline	- Cutbacks in lines - Segment are too short
Table	- Unique ID - Invalid strings value
Topology	- Dangles on line features

	- Orphan topology feature - Unnecessary polygon boundaries - Unnecessary pseudo nodes

3.2.2 Sampling Plan

Section 2.3 has reviewed the sampling methods for evaluating conformance to spatial data. This section presents a sampling plan which specifies the criteria for defining samples and selecting sampling methods. Sampling design requires knowledges of the distribution of homogeneous features across the testing area, the number of samples to be taken, and choice of sampling methods.

In this section, a lot represents the minimum unit for quality evaluation, and an item represents the minimum unit that to be inspected such as a physical feature, an attribute or a domain, etc. The sampling plan contains four steps, which are: a) select items, b) divide items into lots, c) divide lots into sampling unit and d) inspection of items.

The selection of items is based on criteria set forth in the data quality objectives. For examples, if the data quality objective is to check the positional accuracy of building features, all building features within the dataset are selected as items.

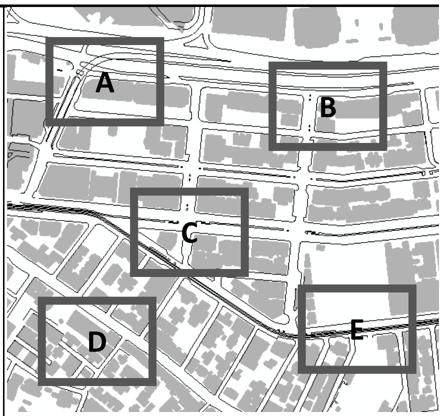
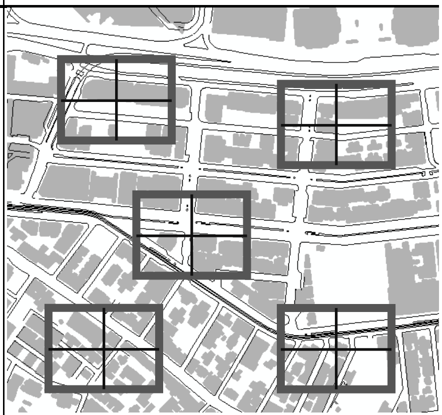
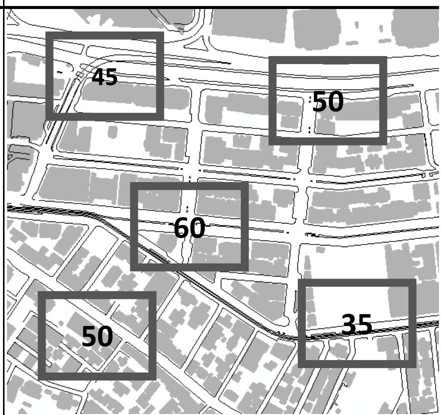
Then, the items are divided into lots. The lots are the minimum unit where the items with homogeneous characteristics are groups into subsets for quality evaluation. The homogeneity can be deduced by the source data of production, production process, or data conversion methods, etc. Using case no.1 in table 3.2 as an example, if the building features are produced by five methods a) ground survey, b) digitization, c) photogrammetry, d) LiDAR, e) conversion from building plans, it is convenience to divide the buildings into five subsets based on the data production methods. Since each subset is assumed to achieve homogeneity in terms of quality, the smaller lot size can be achieved.

Area-guided sampling method can be used to divide lots into sampling units which the inspection is conducted. The total numbers of items within the selected sampling units should be specified in accordance with the minimum criteria of quality evaluation. Area-guided sampling can be used as the first stage of sampling, and then feature-guided sampling can be used to select items in each unit.

In the final step, the sample size is defined and all items which belong to the selected sampling units are inspected. Table3.4 shows an example of sampling plan.

Table 3.4 An example of sampling plan for assessing the positional accuracy of building features in a dataset

Process Steps	**Descriptions**	**Examples**
Select items	All buildings within the dataset are selected as items	

Divide items into lots	Since the buildings are captured using five methods, the buildings are divided into five lots (polygons “A”, “B”, “C”, “D”, and “E”) for quality evaluation.	
Divide lots into sampling unit	Each lots are divided into sampling units based on area-guided sampling methods.	
Inspection of items	Define sampling size (no. of buildings, “45”, “50”, “60” and “35”, will be inspected within sampling units) and inspect each items within the sampling units.	

3.3 Step 3: Determine Data Quality Result

The data quality result can be classified into counting-related data quality result and uncertainty-related quality result. The data quality result should be easily comparable with each other and presented by commonly understanding methods. The data providers should choose a suitable data quality value type to report quality results. Table 3.5 shows three examples of data quality results.

Table 3.5 Examples of data quality result

Example no.	**Data Quality Objectives**	**Data Quality Elements**	**Data Quality Subelements**	**Data Quality Value Type**	**Data Quality Results**
1	The absolute horizontal accuracy of building features	Positional accuracy	Absolute accuracy	Measure	RMSE of distance of building corners is 0.6 m at 95% confidence level
2	All buildings classified as "Police Station" shall bear building address	Completeness	Commission / Omission	Integer	135 (omission)
3	For adjoining 3D RoadPolygons, the vertices with same x, y coordinates should have same z coordinate	Logical consistency	Connectivity error *(new quality subelements)*	Correct items rate in percentage	97.3 %

3.3.1 Counting-related Data Quality Result

The counting-related data quality results are deal with the concept of counting errors. The number of errors can be used to construct different kinds of data quality results such as error rate, integer, and Boolean, etc. The result is determined based on the concept of counting errors or counting correct items. Table 3.6 shows examples of counting-related data quality result.

Table 3.6 Examples of counting-related data quality result

Data quality measure name	**Data quality measure definition**	**Example**	**Data quality value type**
Error indicator	Indicator that an item is in error	False	Boolean (if the value is true the item is not correct)
Correctness indicator	Indicator that an item is not in error	True	Boolean (if the value is true the item is correct)
Error count	Total number of items that are subject to an error of specified type	11	Integer
Correct items count	Total number of items that are free of errors of a specified type	571	Integer
Error rate	Number of the erroneous items with respect to the total number of items that should have been present	0.0152, 1.52%, 1:285	Error rate can either be presented as real, percentage or as ratio

Correct items rate	Number of the correct items with respect to the total number of items that should have been present	0.962, 96.2%, 135:160	Correct items rate can either be presented as real, percentage or as ratio

3.3.2 Uncertainty-related Data Quality Result

The uncertainty-related data quality results are based on the concept of modelling the uncertainty of measurements with statistical methods. Numerical values are obtained from some kinds of measuring procedure can only be observed to a certain accuracy. Assuming the measured quantity is random variable, the uncertainty can be quantified. The methods used to defined uncertainty-related data quality result are based on three assumptions: a) uncertainties are homogeneous for all observed value, b) the observed values are not correlated, and c) the observed values have normal distribution.

The true value is never known, the uncertainty-related data quality result give the probability of the true value to be within a certain interval. This interval is called the confidence interval or significance level. It is given by the probability of the true value begin between the lower and upper limit. Section 2 has fruitfully reviewed the relevant theories.

3.4 Step 4: Report Data Quality Result

The common ways currently used to report data quality result are metadata and quality report. Metadata data provides a common structure for data quality descriptions. Metadata is data about data, and provides other information such as sources, spatial extent, general content, production process and responsibilities.

Data quality report is used when the data quality result reported as metadata is only pass/fail. The data quality report also uses to explain and interpret the meanings of the aggregate quality result.

3.4.1 Report as Metadata

Metadata is developed using Unified Modeling Language (UML) which provides a robust, object-oriented structure that assists to visualize complex relations among sections, the information contained with the sections, as well as the information from related standards. Metadata is now the most popular method for describing digital geographic data. Metadata is applicable to independent datasets, aggregation of dataset, individual geographic features, and the various classes of objects that include a feature. ISO 19115 (2003) presents UML models for data quality information, lineage information, and data quality classes and subclasses, which are shown in figure 3.8 (a), (b) & (c).

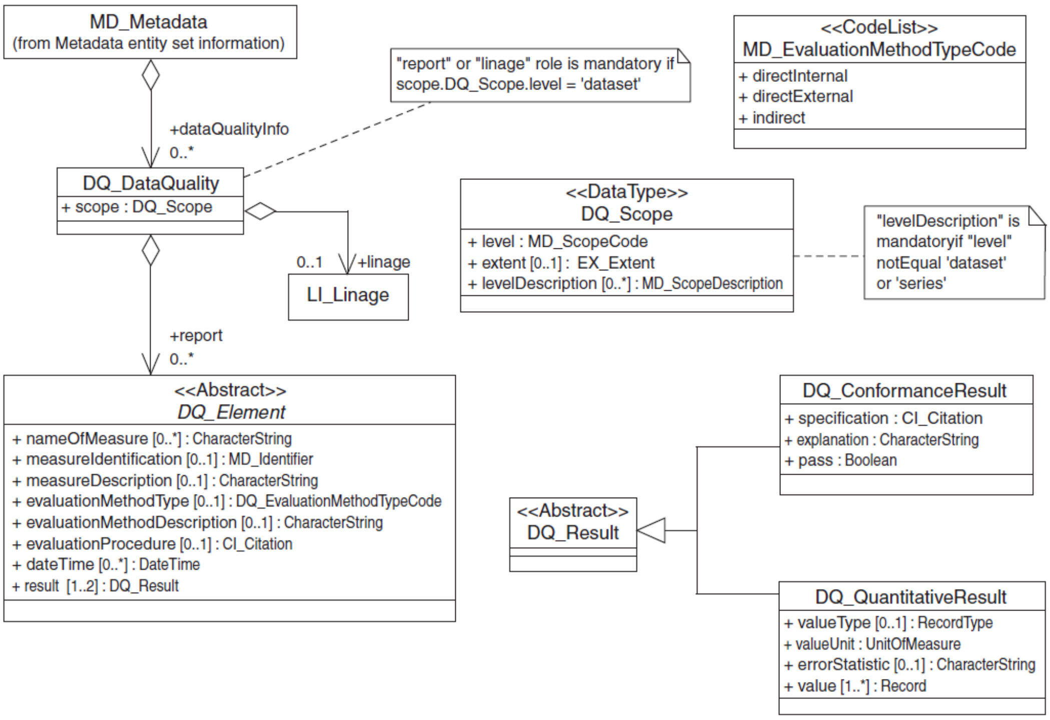


Figure 3.8 (a) Data quality information (ISO 19115, 2003)

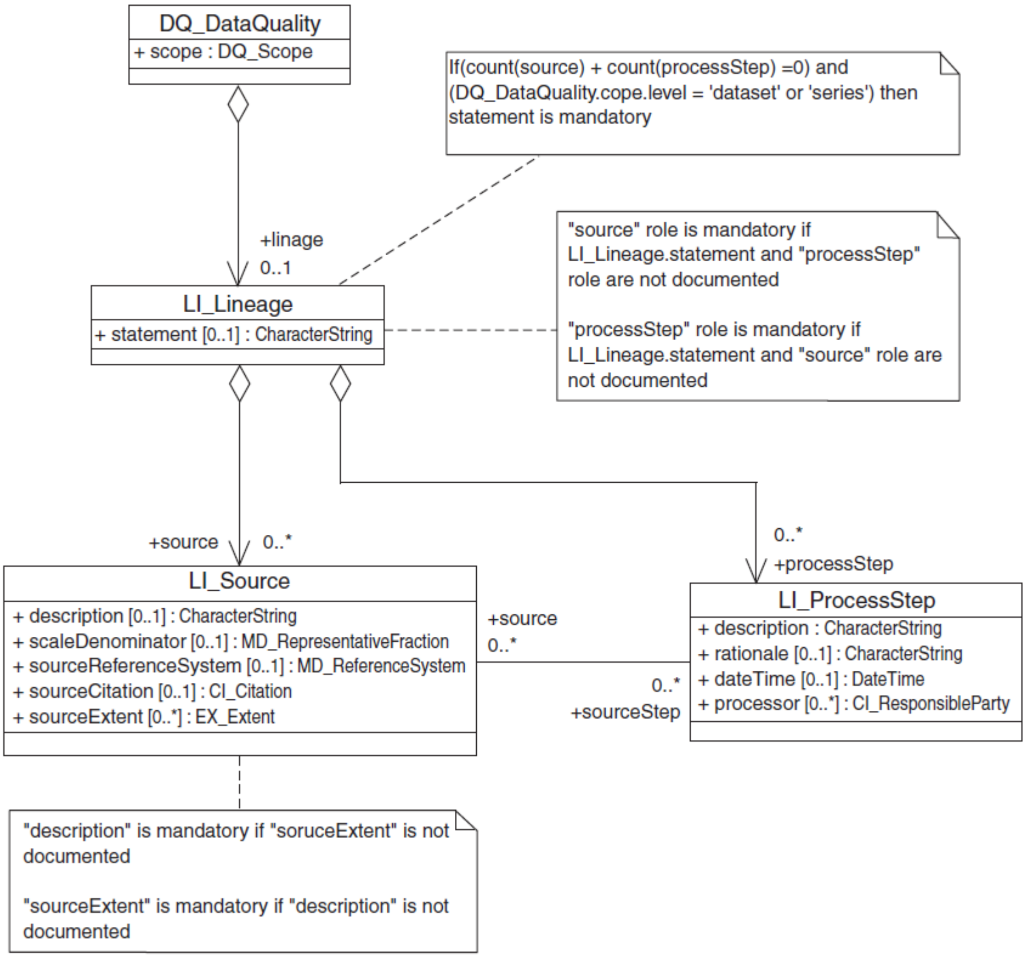


Figure 3.8 (b) Lineage information (ISO 19115, 2003)

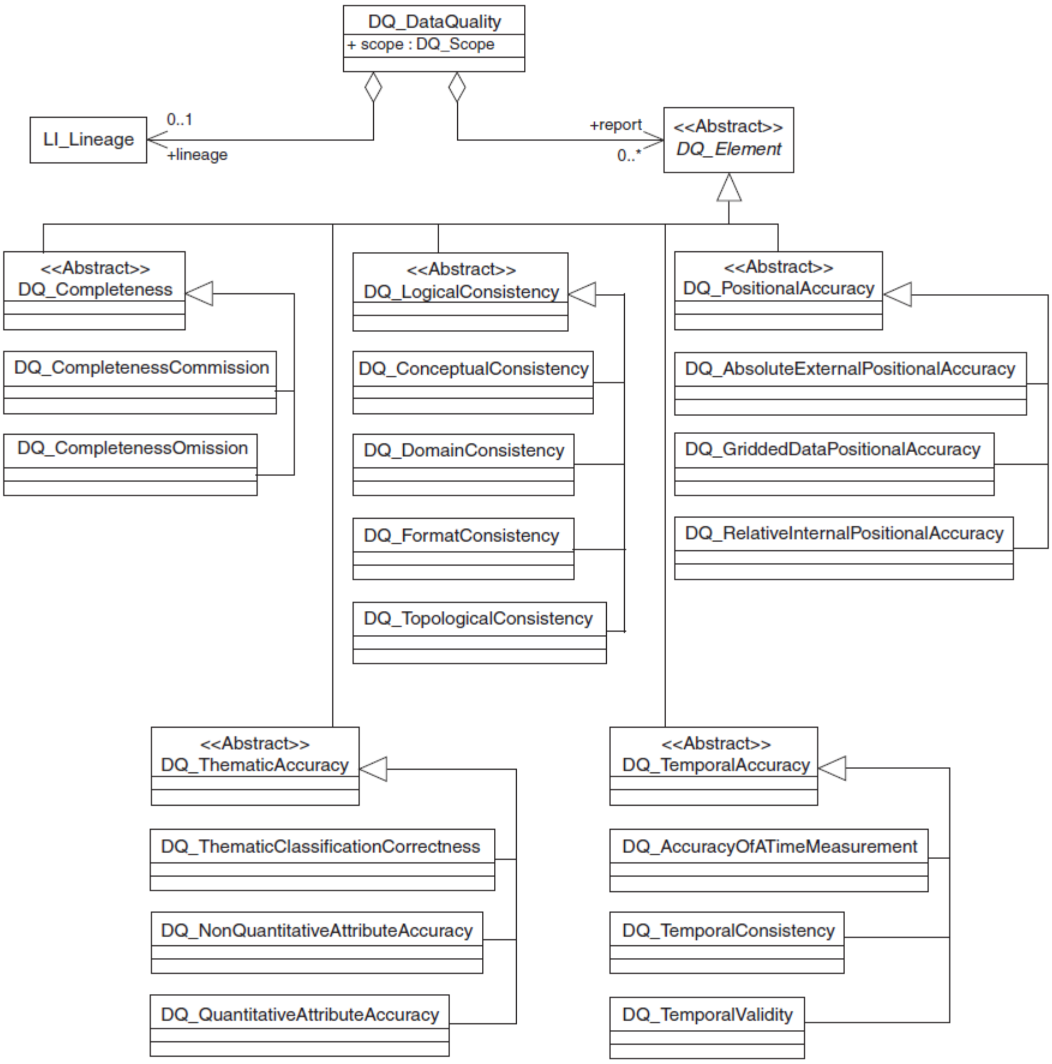


Figure 3.8 (c) Data quality classes and subclasses (ISO 19115, 2003)

3.4.2 Report as Data Quality Report

A data quality report shall include at least the data quality objectives, data quality evaluation methods, sampling methods, data quality result and other general information such as evaluation date, etc. The combination of table 3.1, 3.2 and 3.3 can be treated as an example of data quality report.

Sometimes, the quality of a dataset is reported by aggregated data quality results. This combines the quality results from difference tests. For example, the quality of a dataset is tested using three quality subelements: absolute vertical accuracy, domain consistency and commission error. Each data quality result is computed and given a Boolean value (v) of “1” if it passed and “0” if it failed. The aggregated data quality result (AQR) is determined by $AQR = v1 * v2 * v3$. If the $ADR = 1$, the overall data quality result is pass. If $AQR = 0$, the result is fail.

CHAPTER 4

CASE STUDY: A REVIEW OF HONG KONG 1:1,000 i-Series TOPO MAP

The Survey and Mapping Office (SMO) of the Hong Kong Special Administrative Region (HKSAR) is a Government agency providing land surveys and all types of mapping services and products for both public and private sectors in Hong Kong. In 2011, SMO launched a Land Information System (LIS) which replaced the Computerized Land Information System (CLIS) serving for nearly 20 years. LIS is a centralized database with seamless data supporting rapid and intensive developments of Geographic Information System (GIS) in Hong Kong.

In 2014, SMO started the Hong Kong i-Series Digital Maps services which introduced four new products: iB1000, iC1000, iG1000 and iB5000. iB1000 and iB5000 are digital topographic maps with scale 1:1,000 and 1:5,000 respectively. iC1000 is a digital land boundary map and iG1000 is a geo-reference database storing attribute information, such as building names, addresses and site names, etc., and corresponding graphical features. The i-Series Digital Maps provide five data formats for different applications:

a) File geodatabase (v9.3.1),
b) Geography Markup Language (GML, v3.1.1),
c) Micro-station (DGN, v8),
d) AutoCAD (DWG, v2000) and
e) Geo-TIFF

This chapter reviews the Hong Kong 1:1,000 i-Series TOPO MAP, hereafter named iB1000, in terms of data quality management and communication of data quality information.

4.1 Introduction

iB1000 maps the city of Hong Kong which facilitating town planning and development, online and mobile mapping services, and other GIS related applications. SMO is committed to provide the most efficient and effective survey and mapping services and products to meet users' requirements.

iB1000 is a vector-based topographic data which represents real-world objects in point, line and polygon, and each with unique reference, named FeatureID. Three sets of attribute information, address, building name and site name, with cross-reference to geometry features, are stored in iB1000. iB1000 have a hierarchical structure. The highest level of the structure is the map layer. Within each map layer is a series of datasets. Within each dataset, there are feature classes. Within each feature classes, there are features representing the real world objects. The dataset groups several feature classes together to provide high-level means of dividing the data logically in the map layer. Text and symbol are used to provide additional information and context about the real world objects. They stored in annotation feature classes which indicate the locations where the text and symbol are displayed.

iB1000 contains seven datasets, and they are: Buildings, Transportation, Hydrography, PlacesofInterest, Relief, Utilities and LandCover. iB1000 is designed to take advantages of the latest technologies in spatial data management. It combines geometry and attribute information that can be more readily manipulated and searched by spatial query and classification tools.

4.1.1 History of iB1000

In the early 1950s, one set of map sheets covered New Territories of Hong Kong was produced at scale 1:1,200. They aimed to supplement the old Demarcation District Sheets produced in the early 1900s. In 1960s, it was realized that an accurate and large-scale maps with contours were required for the purposes of town planning and city development. In 1962, a British company was awarded to carry out an aerial survey for mapping purpose. This project was accomplished in 1971, which produced a map series at scale 1:600, five-foot contours and covered Hong Kong Island, Kowloon and New Kowloon. The remaining areas and New Territories were mapped at scale 1:1,200, with 10-foot contours.

In 1970s, SMO converted some 3,000 sheets to the metric scale of 1:1,000 with metric contours and spot height. This work was greatly assisted by the establishment of Photogrammetric and Air Survey Section in 1976. To complete a full coverage of the map series, some 150 additional sheets were produced by mapping program to survey the unmapped areas by photogrammetric method. This project was started in 1994 and completed in early 1997. Up to now, iB1000 contains about 3,280 survey sheets at 1:1,000 scale. Updating of iB1000 is carried out continuously by SMO using diverse data capture methods.

4.1.2 Data Capture Methods

iB1000 is produced using five types of data capture methods, and they are: digitization, photogrammetry, field survey, Light Detection and Ranging (LiDAR) survey and external sources.

At present, map digitization is seldom used by SMO. Converting the data from as-built engineering plans and building plans (i.e. external sources) is one of data sources to create major infrastructures, such as building estate, highway, bridge and large public area. Photogrammetry is the main data source of iB1000. Field surveys, including, tile survey, topographic survey, GPS, precise levelling, ordinary levelling, trigonometrical height, are used for partial updating. Starting from 2012, building base levels are acquired for Digital Terrain Model generated from LiDAR data.

Table 4.1 shows the data capture methods of buildings and roads in iB1000. It indicates that about 70% of buildings are captured by digitization, and 57% and 27 % of roads are captured by digitization and topographic survey respectively. Table 4.2 shows the data capture methods of spot height in iB1000. It shows that over 79% of spot heights are captured by photogrammetry.

Table 4.1 Data capture methods of buildings and roads

Data Capture Methods	**Buildings**		**Roads**	
	No. of Records (total: 245,565)	**%**	**No. of Records (total: 111,051)**	**%**
Title Survey	11,930	4.85	1,206	1.08
Topographic Survey	46,314	18.86	29,507	26.57
Topographic Survey by GPS	3,410	1.39	304	2.74
1:1000 Photogrammetric Plot	9,205	3.75	12,723	11.46

1:1000 Survey Sheet (converted by SMO)	60,652	24.70	34,257	30.85
1:1000 Survey Sheet (converted by contractor)	112,255	45.71	29,371	26.45
1:1200 Survey Sheet	1,230	0.5	258	0.23
1:5000 Survey sheet	3	0	29	0.3
Estimation / Proposal	420	0.17	429	0.39
Building Plan	129	0.05	213	0.19
Mixed	12	0	16	0.01
Unknown	5	0	0	0

Table 4.2 Data capture methods of Spot Heights

Data Capture Methods	**Spot Height**	
	No. of Records (total: 347,207)	**%**
Precise Levelling	247	0.07
Ordinary Levelling	19,357	5.58
Trigonometrical Height or GPS	12,019	3.46
Photogrammetry	277,092	79.81
External Sources	35,200	10.14
3D Project	26	0
Trigonometrical Height only	466	0.13
GPS only	1,656	0.48

Estimation (e.g. by no. of storey)	942	0.27
Building Plan	202	0.06
Unknown	0	0
LiDAR	0	0

At present, the precision code which represents the data capture methods is used to denote the positional accuracy of features stored in iB1000. Based on the law of error propagation under the assumption that error components follow normal distribution and uncorrelated, the horizontal accuracy of these data capture methods can be determined. For example, the error in title survey, E_t, can be measured by:

$$E_t^{\,2} = e_r^{\,2} + e_c^{\,2} \qquad (4.1)$$

where e_r = Error of radiation

e_c = Error of ground control stations

According to equation (4.1), if the positional error of radiation e_r for house lots and agricultural lots is assumed to be 2.4 mm and 11.8 mm respectively (Chan K.K. 1994), the minimum accuracy for tile traverse is 1:7,500, and the average distance between ground control for tile survey is 100m,

e_c = *100 / 7500 = 13.3* mm

Then,

$E_t = \sqrt{(2.4^2 + 13.3^2)} = 13.5\ mm$ (at the best case)

$E_t = \sqrt{(11.8^2 + 13.3^2)} = 17.8\ mm$ (at the worst case)

Based on this assumption, table 4.3 shows the estimated horizontal accuracy of several data capture methods of iB1000. The weakness of using precision code to denote the positional accuracy of iB1000 is discussed later in Section 4.4.

Table 4.3 Estimated horizontal accuracy for several data capture methods of iB1000

Data Capture Methods	**Collection Methods**	**Positional Accuracy**
Title Survey	Ground Survey	13.5 mm - 17.8 mm
Topographic Survey	Radiation	11 cm
	GPS	11 cm
	Tie Measurement	8 cm – 17 cm
1:1000 machine Plot	Photogrammetry	0.65 m
1:1000 Survey Sheet (converted by SMO)	Digitization	0.4 m – 2.4 m
1:1200 Survey Sheet	Digitization	0.5 m – 2.9 m
1:5000 Survey sheet	Digitization	N/A

4.1.3 Data Dissemination

The Hong Kong Map Service was launched in 2010, which offered customers a comprehensive range of digital map products and the related data dictionaries. This is a web-based ordering system that allows the customers to order iB1000 and other map products. Customers are allowed to order iB1000 in two ways: a) order of discrete tile digital map and b) order of seamless digital map. For discrete tile digital map, the deliverable is sheet-based map according to the user-selected sheet number(s). For seamless digital map, the deliverable is one single file of digital map in seamless nature. A standard map sheet covered ground areas 750 m x 600 m. Figure 4.1 shows an example of iB1000 (Geo TIFF) survey sheet. Figure 4.2 shows the interface of Hong Kong Map Service.

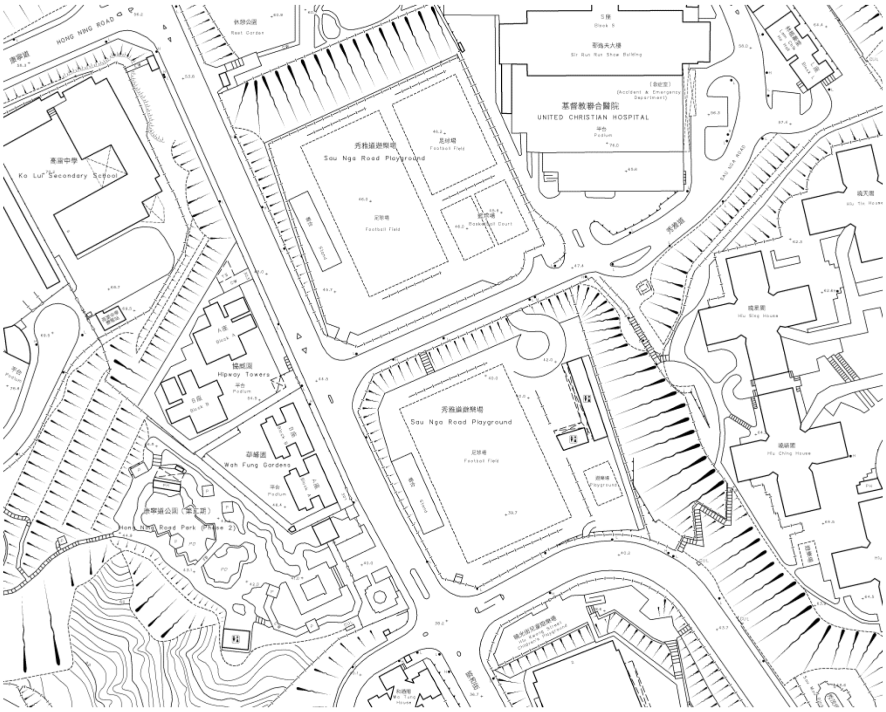


Figure 4.1 An example of iB1000 (Geo TIFF) survey sheet

Available at:

https://www2.hkmapservice.gov.hk/DDS/static/html/CLIS/iSeries_public_eng.htm

[Accessed 3 May, 2016]

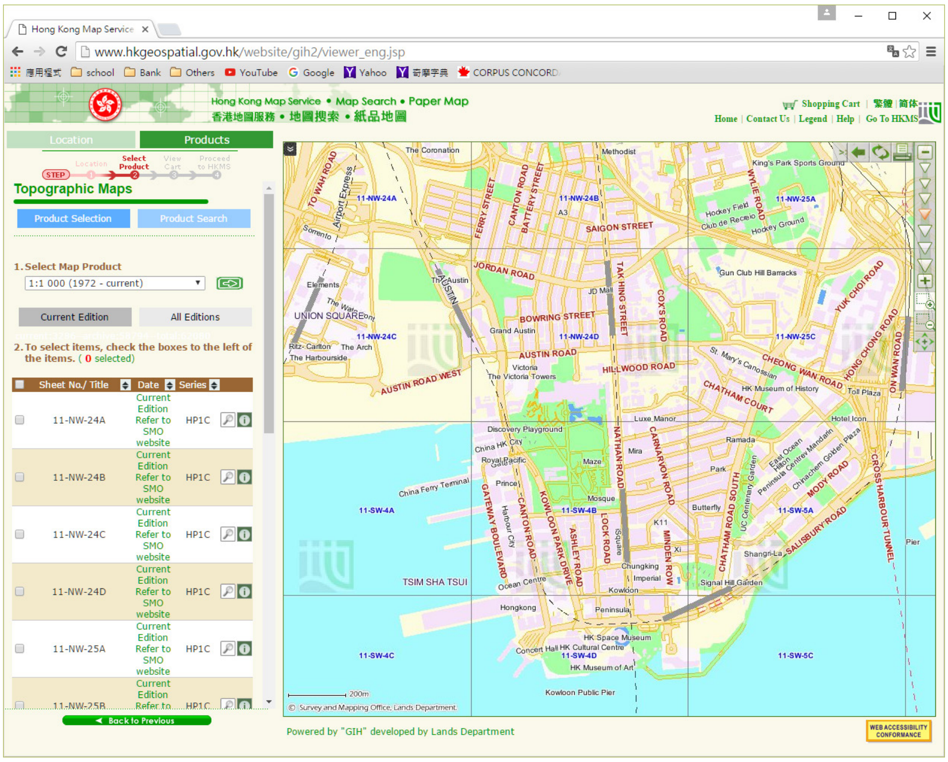


Figure 4.2 Interface of Hong Kong Map Service

Available at: http://www.hkgeospatial.gov.hk/website/gih2/viewer_eng.jsp

[Accessed 3 May, 2016]

4.2 Data Quality Management of Hong Kong 1:1,000 i-Series TOPO MAP

Spatial data is a long-term asset which requires continuous update and data management to maintain the quality. Product standards and data quality control are two major factors for good data quality management. This section reviews the product standards and data quality control of iB1000.

4.2.1 Product Standards

"1:1,000 Basic Mapping Specifications (version 4.1), 2015" and "District Survey Office Technical Manual (version 4.2), 2014" are documents describing the definitions of spatial features and providing standards for data capture methods of iB1000. The rapid development on GIS technologies and applications driving SMO continuously reviewed these documents to fulfil customers' requirements.

4.2.1.1 1:1,000 Basic Mapping Specifications (version 4.1)

1:1,000 Basic Mapping Specifications (version 4.1), hereafter named Mapping Specifications, provides comprehensive guidelines on how to update the iB1000 in the Land Information System (LIS). It includes the definition of map features and their attributes, and the cartographic representation of these features on the iB1000 standard survey sheets produced by the SMO.

According to the Mapping Specifications, the general principle of data capture is to record "what is seen in the real world". This document further introduces the definitions of each ground features and their attributes, and the generalization rules on feature representation.

Mapping Specifications also includes the revision policy of major infrastructure, for examples, main roads, prominent buildings, large site formation works and large public utility installation, etc. Table 4.4 lists the major infrastructures together with an assigned "Date Code" that indicates the appropriate time for survey commencement.

Table 4.4 Revision policy of major infrastructures

Type of Infrastructures	**Date Code**
Expressway, highway, bridge, tunnel	1
Major road (longer than 0.5 km)	1
Public housing estate	2
Government building	2
Public park and garden, amusement park	3
Large site formation (larger than 3 ha in area)	3

Definitions of Date Code:

Code 1 – Upon completion or road surfacing

Code 2 – Upon demolition of scaffolding/hoarding

Code 3 – Upon substantial completion of construction work

Mapping specifications also contains the horizontal and vertical accuracy of four data capture methods, table 4.5 lists the horizontal and vertical accuracy of these data capture methods.

Table 4.5 Horizontal and vertical accuracy of four data capture methods

Data Capture Methods	**Horizontal (±)**	**Vertical (±)**
Cadastral Survey	0.1 m	N/A
Topographic Survey	0.2 m	0.1 m
Machine plots produced by Photogrammetric Unit	0.3 m	0.4 m
Old paper map	1 m	N/A

iB1000 defines more than 100 types of features in real world. To cater for the increasing demand of digital topographic data and the rapid development of various GIS applications, SMO continuously keeps track of the user's requirements and revised the Mapping Specifications. The data management team of SMO is responsible to ensure the data fulfills the requirements stated in Mapping Specifications and conducts assessments to measure the data quality. Table 4.6, 4.7 and 4.8 show three examples of data quality assessments performed by SMO to ensure the quality of building, road and lamp post in iB1000.

Table 4.6 Quality assessment of building

Statements in Mapping Specifications	**Data Quality Check for Acceptance**
All permanent buildings or structures of a size larger than 4 m^2 should be surveyed. A smaller building forming a prominent landmark may be surveyed and shown for identification purpose.	Check the completeness of building: - If a building less than 4 m^2 are presented in iB1000, verify whether it is prominent landmark. If not, it is considered as commission error
Building top level, if collected through ground survey or other means, should be recorded in the attribute field of "Roof Level". The building top level of a building is defined as the highest level of the largest accessible area on the rooftop of a building, excluding water tank and lift shaft. In case	Check the vertical accuracy of building top level based on the definition: - Conduct vertical accuracy on building top level

there is no largest area on the rooftop, the building top level is the level at the highest point or the ridge of the building..	
Input the building name or block number into the table	Check the completeness of building name: - If there is a building name does not inputted in the table, it is an omission error

Table 4.7 Quality assessment of road

Statements in Mapping Specifications	**Data Quality Check for Acceptance**
In general, all road margins are to be surveyed and shown to define road alignment. If traffic islands and central dividers are considered to be significant, they are treat as road margin and surveyed.	Check the horizontal accuracy of road margin on well-defined nodes or vertices: - Conduct horizontal accuracy on road margin
Annotation of street name, gazette or ungazette, should be shown in upper case.	Check the logical consistency for the annotation: - If gazette or ungazette street names do not show in upper case, it is a format consistency error

Table 4.8 Quality assessment of lamp post

Statements in Mapping Specifications	Data Quality Check for Acceptance
Survey the lamp pole erected along the roadside, on the street or in the road-divider, etc.	Check the completeness of lamp pole within specific area: - If there is a lamp pole, erected along the roadside, on the street or in the road-divider, does not inputted in iB1000, it is an omission error
Lamp post may have one or more numbers, record than as separate records in attribute table "UtilityNumber".	Check the completeness of utility number for each lamp post: - If there is a lamp post number does not inputted in the table, it is an omission error

4.2.1.2 District Survey Office Technical Manual (version 4.2)

District Survey Office Technical Manual (version 4.2), hereafter named Technical Manual, mainly provides guidelines on Land Boundary Survey, for examples, the actions on new land grants and allocations, re-establishment of land boundary, land acquisition and land exchange, etc. Chapter 11 of Technical Manual is related to topographic survey which contains the data capture standards and guidelines for updating iB1000. The followings are revision policy, horizontal control standards and vertical control standards for iB1000.

Revision Policy:

The Technical Manual states that the major infrastructures would be surveyed and shown on iB1000 within 12 weeks after completion. For other features, SMO collects data regularly from various sources and adopts the continuous updating approach to update the iB1000 and carries out the updating work on a job basis.

Horizontal Control Standards:

Technical Manual defines the accuracy requirements for establishing horizontal control network. These requirements are applied to field survey and GPS survey, and they are:

Traverse:

Check the origin of bearing and coordinates by observing to at least two known control survey stations and measuring a distance to at least one of the known control survey stations. All closing stations and bearings of a traverse should be properly checked. Hanging and self-closed traverses are not allowed. However, under exceptional circumstances, such as site constrain factor, self-closed traverse may be permitted with prior approval of unit/section head. Table 4.9 shows the required accuracy of a traverse for topographic survey.

Table 4.9 Required accuracy of a traverse for topographic survey

Type of Misclosure	Requirements
Angular misclosure	$40\sqrt{n}$ seconds where: n is the number of survey stations in the traverse
Linear misclosure	$10 + S/4$ millimetres where: S is the total length of the traverse in metres

GPS:

The requirements for static or rapid static GPS survey are as follows:

a) The number of satellites observed at any one time shall not be less than 5,

b) The geometric dilution of precision (GDOP) value during observation shall not be greater than 5,

c) The elevation mask for satellite observation shall not be less than 15 degrees above the horizon,

d) The epoch recording rate shall be a multiple of 5 seconds, and

e) The span of the observation period should be at least 15 minutes to allow for an iono-free fixed solution to be possible.

Vertical Control Standards:

The vertical control standards for ordinary levelling and height traversing are defined as below:

Ordinary Levelling:

The requirements for ordinary levelling are as follows:

a) Start a levelling route from and close onto different points of known levels. These points are either bench marks established by the Geodetic Survey Section or temporary bench marks established in previous survey,

b) Carry out a two-peg test before survey. The collimation error shall not exceed 3 mm/50 m (i.e. 12” of arc) or as specified in the equipment user manual,

c) Keep the foresight and backsight distance approximately equal in each setup,

d) Ensure that the line of sight is 0.3 m above ground, and

e) The misclosure for an ordinary levelling shall not exceed $0.012\sqrt{K}$ metres, where K is the distance levelled in kilometres.

Height Traversing and Reciprocal Height Observation:

The requirements for height traversing and reciprocal height observation are as follows:

a) Measure the height of instrument and target at each setup,

b) Observe the forward and backward vertical angles from instrument to targets in both FL and FR positions,

c) The allowable misclosure of a height traversing is $10 + 10\sqrt{n}$ millimetres, where n is the number of traverse legs involved, and

d) Carry out height traversing by reciprocal height observation.

iB1000 is produced using diverse data capture methods, some features, converted by old paper maps, may display consistent discrepancies with newly created features. If the discrepancies are greater than 0.5 metre, it is considered as “Block Shift” and should be update using new survey methods.

4.2.2 Data Quality Control

With the support of SMO's management, a procedure for continuous updating and related quality checking system have been established. The procedure is introduced below.

4.2.2.1 Procedure for Continuous Updating

Operation Procedures (SMOP) is a set internal documents used by SMO's staff for various operational needs. There is a procedure for data collection and continuous updating of the iB1000. Figure 4.3 shows the summary of this operational procedure.

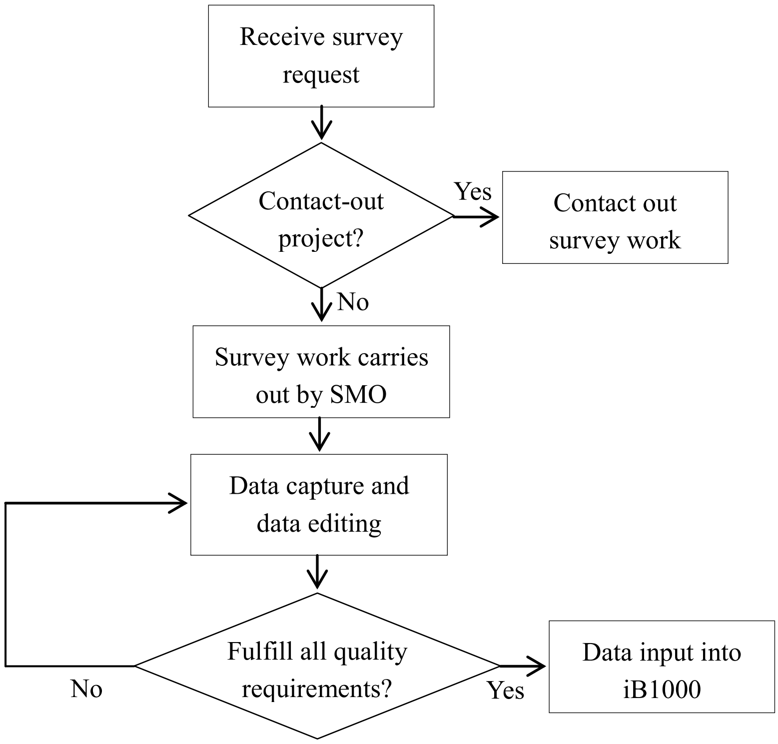


Figure 4.3 Procedure for data collection and continuous updating of iB1000

4.2.2.2 Quality Control of 1:1,000 i-Series TOPO MAP

The quality control framework for iB1000 consists of two major iterative processes which involve the inputs from front-end users and back-end data management team.

For the front-end users, data input and updates are performed by manual process. There are data production tools and quality checking tools streamlining the workflow and reviewing the quality of newly-created features. After completing the front-end checking processes, the front-end users input the data to the centralized database of LIS.

The back-end data management team will conduct checking on the master database in monthly basis so as to satisfy the level of fulfilment of the quality standards. All features, their relationships and attributes will be checked by in-house programs. Once an error is identified, the back-end data management team will decide the appropriate problem fixing actions. In some cases, field verifications are required to verify the correctness of the captured data. Figure 4.4 below shows the quality controlling framework of iB1000.

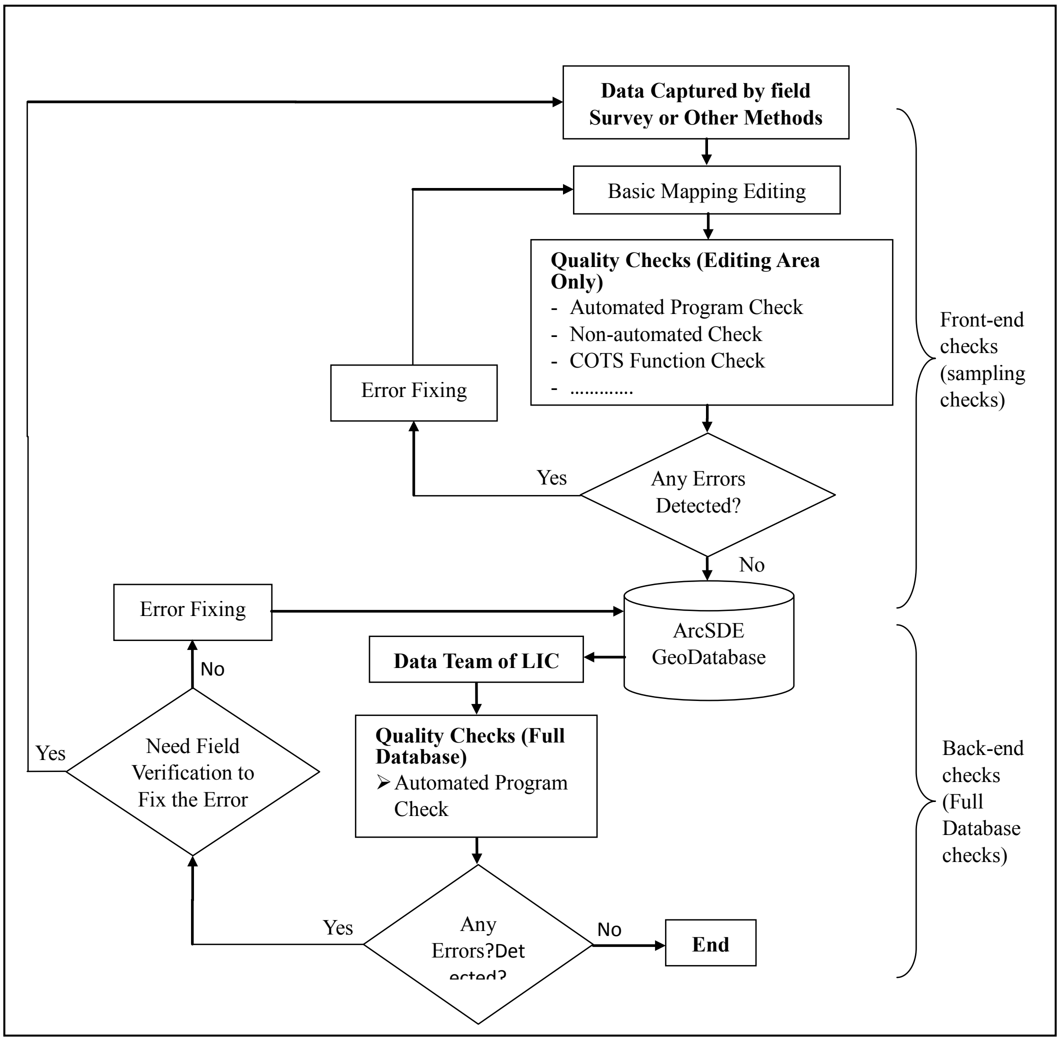
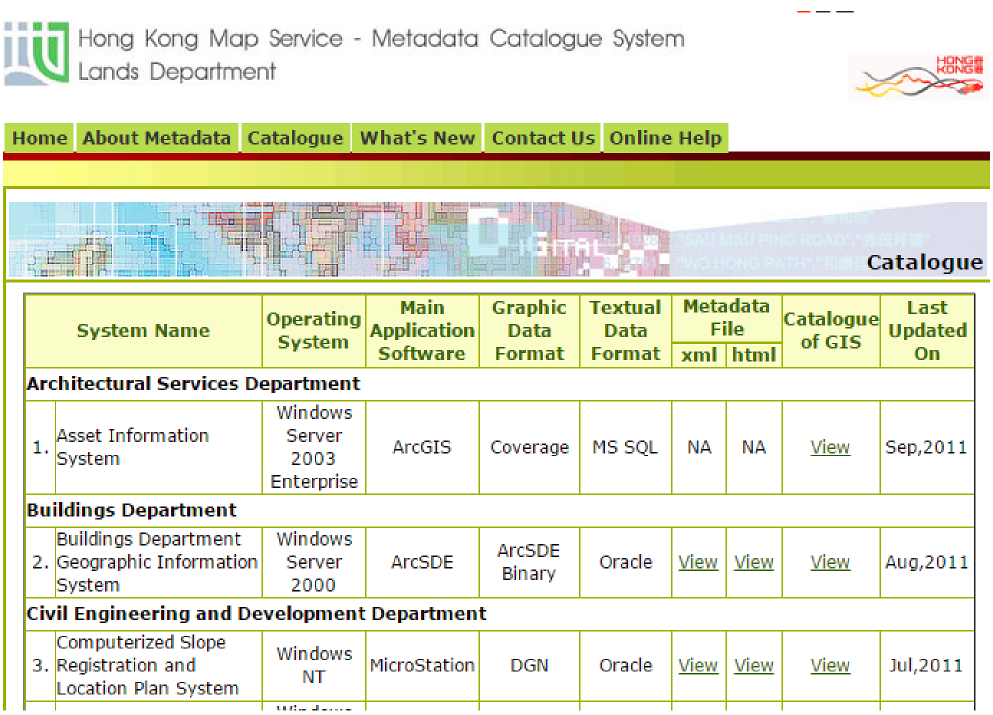


Figure 4.4 Quality control framework of iB1000

4.3 Provision of Data Quality Information

The quality information of iB1000 can be obtained from Hong Kong Map Service – Metadata Catalogue System. This system is a repository of the metadata of the Digital Geospatial Data (DGD) owned by Government Departments that took part in a data sharing program named Data Alignment Measures (DAM). The Metadata Catalogue System provides web access functions for viewing the metadata online. All metadata provided are conforming to the US Federal Geographic Data Committee (FGDC) for quick reference for data information. The quality information, attribute accuracy, positional accuracy, completeness, of iB1000 is quoted below for reference. Table 4.5

shows the interface of the metadata catalogue.


Hong Kong Map Service - Metadata Catalogue System
Lands Department

Home | About Metadata | Catalogue | What's New | Contact Us | Online Help

Catalogue

	System Name	Operating System	Main Application Software	Graphic Data Format	Textual Data Format	Metadata File		Catalogue of GIS	Last Updated On
						xml	html		
Architectural Services Department									
1.	Asset Information System	Windows Server 2003 Enterprise	ArcGIS	Coverage	MS SQL	NA	NA	View	Sep,2011
Buildings Department									
2.	Buildings Department Geographic Information System	Windows Server 2000	ArcSDE	ArcSDE Binary	Oracle	View	View	View	Aug,2011
Civil Engineering and Development Department									
3.	Computerized Slope Registration and Location Plan System	Windows NT	MicroStation	DGN	Oracle	View	View	View	Jul,2011

Figure 4.5 Hong Kong Map Service - Metadata Catalogue System

Available at: http://www.hkmapmeta.gov.hk/mcs/home/en/home.htm

[Accessed 3 May, 2016]

The data quality information of iB1000, which stored in metadata catalogue system, is as below:

Attribute_Accuracy:

Attribute_Accuracy_Report:

Attribute accuracy is checked by manual comparison of the source with the hard copy print-out. The attributes items are also confined by the Item Definition and pre-defined feature type. In addition, DM software tests the attributes against a master set of valid attributes. The program will prohibit *data writing to the data library and prompt a warning message if any deviation from the standard is found.*

Logical_Consistency_Report: Polygon and chain node topology present

Completeness_Report:

In the 1:1000 scale BMS dataset, only those features defined in the B1000 specifications will be captured. If two or more ground features are separated by less than 1 metre, only the most prominent feature will be shown whereas the others will be suppressed in the graphic presentation. Besides, features may have been eliminated or generalized on the source graphic, due to scale, legibility or change of mapping specifications. Owing to stringent time-frame during data conversion, some of the attribute fields are yet to be filled up.

Positional_Accuracy:

Horizontal_Positional_Accuracy:

Horizontal_Positional_Accuracy_Report:

Features in BMS have a "Precision" code to denote its data capturing method or sources. The code ranges from 1 to 9 with 1 being the most accurate and 9 being the least. Accuracy analysis has been carried out to quantify how each of the ranking represents. The analysis is based on the law of propagation of error with the assumption that errors are uncorrelated and normally distributed. The analysis shows that the position accuracy of BMS data varies from about 0.015m in the best to 1.500m to the worst for data captured by title survey method to data digitized from the original 1:1200 scale survey sheet.

Process_Step:

Process_Description:

All the Chinese text is added to the database as graphic and textual data

Process_Date: 199907

Process_Step:

Process_Description:

Three proposd layer; PBLDGPOLY, PROPROAD and PROPINFRA are added to B1000 Mapping System

Process_Date: 200108

Process_Step:

Process_Description:

To meet the requirements of DAM (Data Alignment Measures), numbers of enhancement and modification is taken into 1:1000 Basic Mapping System. New layers (DBLDGPOLY, PODPOLY, PPODPOLY and DPODPOLY) and lookup tables (PODIUM_NAME and BLDG_CSUID) are added. Numbers of the existing layers (BLDGPOLY and BLDG) and lookup tables (ADDR_HSNO, BLDG_NAME and GEO_REFERENCE) are modified. The layer OBLDG is removed as it is replaced by DBLDGPOLY. Also, some of the modification are not related to DAM exercise; new layer PSPOTHT is added and existing layer SPOTHT is modified.

Process_Date: 200505

4.4 Conclusions

This chapter has thoroughly reviewed the characteristics of iB1000 including the data capture methods, data control framework and data management system. The quality control mechanism has long been established in SMO which covers entire product lifecycle. Clear working procedures and product specifications provide guidelines for continuous updating. With the implementation of LIS, i-Series digital spatial data is now available with different format supporting various GIS applications and analysis in Hong Kong. Although a quality management system has been developed, there is still a room for improvement. The followings are some quality problems in iB1000, the next section will provide corresponding solutions.

a) There is no data quality standard defines the requirements in data capture, data updating and data conversion. For example, the threshold of horizontal and vertical accuracy requirements of iB1000 is not defined yet,
b) There is no systematic data quality assessment procedure for continuous updating. For example, a building plan is used for updating, the acceptance test should encompass clear data quality information which describes the positional accuracy, completeness, and logical consistency of the building plan and judge whether the building plan is acceptable for iB1000 updating,
c) Lack of data quality information provides to users and reported in metadata,
d) Using precision code to denote the level of horizontal and vertical accuracy is too crude. Because the positional accuracy is determined by deduction methods where some assumptions, the distance of radiation is 100 m, may not be applicable during data field survey and the error in data editing is ignored in the determination,
e) Some precision code is meaningless, for example, unknown,
f) There is no logical way to manage the data quality information. For example, adopting the international standard to assess and report the data quality results.
g) When the requirements in product specifications are changed, there is no mechanism to report the impact on data quality, and
h) Lack of data currency information, for example data updating date, field survey date, and date of source material, such as aerial photos, etc.

CHAPTER 5

CASE STUDY: PROPOSED DATA QUALITY CONTROLLING SYSTEM FOR HONG KONG 1:1,000 i-Series TOPO MAP

The value and usefulness of spatial data is directly related to its quality. From a data provider's point of view, the major objective of data quality control is to ensure the dataset conforms to the data standards, the product specifications and fulfils users' requirements. The data standards define the required quality level of a dataset, for examples, cadastral dataset requires positional accuracy in few centimetres, whereas a countryside map requires a positional accuracy of several metres. The product specifications describe the rules of generalization, the definitions of geometry and attribute of spatial and non-spatial information. The user's requirements are the measures that focus on the "fitness for purpose" of the data to customers' expectations.

SMO is the central authority of land surveys and all types of mapping in Hong Kong. Its responsibilities are to provide land boundary surveys, photogrammetric survey as well as cartographic and reprographic services, and maintain the land information system for mapping data and land boundary records. Continuously improving the quality of services and products are commitments to SMO. In order to achieve this, the development of data quality controlling system is one of the keys. This chapter presents a data quality controlling system for SMO.

5.1 Components of Data Quality Controlling System

The data quality controlling system introduced in this chapter contains five components, they are:

a) Data policy,

b) Data standards,

c) Data quality control,

d) Error handling, and

e) Data dissemination

The details of each component are discussed in this chapter.

5.1.1 Data Policy

Data policy is a high level decision that forms the objectives of a dataset. In establishing or maintaining a GIS, data always plays an important role which closely relates to the existing and potential applications, and cost of the system. Data policy covers wide ranges of areas which encompasses dataset objectives, data acquisition plan, data use and share, and data security. The followings provide general guidelines for establishing data policy of SMO.

Data Objectives:

The data objectives can be derived from the current and potential applications of the system. According to the application requirements, data provider can design the data scale, format, standards, specifications, acquisition methods, revision polices, disseminations and security of the dataset. iB1000 serves both government and private sectors, for example Lands Department, Fire Services Department, Highways Department, Building Department, The Hong Kong Electric Company, and The Hong

Kong and China Gas Company Ltd. etc. Therefore, iB1000 must fulfill both public and private sectors' requirements and design to have common file format, high level of interoperability, good data quality and comprehensive quality information, easily access, good customer service, complete documentations, and fruitful data content, etc.

Data Acquisition Plan:

Data acquisition is an expensive and time consuming task. Mapping agencies and government often adopt multiple data acquisition methods in producing and maintaining spatial data. The aim is to establish a plan that maximizes the data usage and with low maintenance cost. As mentioned in section 1.1.1, spatial data is acquired using either primary or secondary data capture methods. iB1000 is currently adopted various data acquisition methods which including primary data capturing such as field survey and photogrammetry survey, and secondary data capturing such as data conversion of as-built building plan. Prior to conduct data collection, it is necessary to assess whether the data is satisfied the quality requirements stated in product standards and product specifications. As a government mapping agency, SMO also needs to seek possible opportunities to obtain data from other government departments and private sectors. This would greatly improve the spatial data sharing and drive toward to develop a spatial data infrastructure in Hong Kong.

Data Use and Share:

In principle, spatial data is designed to be easily accessible to the public and make available at little or no cost. The data providers shall establish intellectual property rights and license agreements for the users who receive the data. This includes the declarations and conditions of usage and reflects all statutory and non-statutory obligations. iB1000 and other land boundary data are produced to satisfy the both internal and external requirements, so there is some restricted information which is solely used by internal applications. SMO is necessary to develop a data filtering strategy that is able to screen out restricted information which is for internal use only. In addition, there is need to create pricing agreements which are in line with the government policy and demand of the industry.

Data Security:

Data providers shall aware the security issues in disseminating data to different levels of users. SMO serves both internal and external users, there is a need to develop a user account system that allows the dissemination of data in accordance with the level of access granted to the users. The user account system shall be able to cater both internal users and external users and grant access to specific data based on the users' approval level of access.

5.1.2 Data Standards

Data standards allow data users understand the content of data and provide information on spatial and non-spatial components of the dataset. The development of data standards, where possible, shall agree with the format and template used by international standard, national standard, or local standard. The communication standards are likely to be open format, such as HTTP, XML and GML, etc.

Data standards are classified into spatial and non-spatial. Spatial standards describe spatial information, such as map projections, datum, coordinate system, scale, xy domain and accuracy, etc. Non-spatial data describe non-spatial information, such as data capture methods, database structure, data processing standards, data quality information, data usability, data dissemination standards, QA/QC standards, data transfer standards and presentation standards, etc.

SMO is now using the precision code to denote the horizontal accuracy and vertical accuracy, and the data quality standards are unclear. This confuses both data providers and data users. Therefore, it is suggested using the principle of international standards to create product standard for iB1000. An example is as follows.

General

SMO serves as the central authority for the collection and distribution of digital topographic map in Hong Kong. This standard is intended to facilitate the interchange and use of iB1000. iB1000 is collected according to the standards set forth in this document.

Series Descriptions and Usages:

Four distinct types of i-Series are produced by SMO: iB1000, iC1000, iG1000 and iB5000. The current and potential applications of i-Series are to support cartographic representation, spatial data analysis, spatial query, data classifications, data visualization, 3D modelling and decision making, etc.

Data Sources:

iB1000 is derived from various data sources such as, ground survey, photogrammetry survey, digitization, external sources, etc. Their accuracies are contained in product specifications and metadata. The accuracy reflects the ground distance in meters.

Revision Policy:

Major infrastructures would be surveyed and shown on iB1000 within 12 weeks after completion. For other features, SMO collects data regularly from various sources and adopts the continuous updating approach to update the iB1000 and carries out the update work on a job basis.

Data Content:

iB1000 contains seven themes: Buildings, Transportations, Hydrography, Place of Interest, Relief, Utilities and LandCover.

Data Structure:

The structure of i-Series data is described by considering to two subject areas: topology and topological elements.

Topology – the i-Series is based on graph theory in which a 3-dimenionsal diagram is expressed as a graph composed of a set of point, line and polygon in a manner that explicitly expresses logical relationships.

Topological elements – the i-Series is composed of three separates, but related, elements: point, line and polygon.

Extraction Specifications

The data extraction specifications contain all information required for data collection. These specifications tell what is collected as a certain feature, and when and how the features are collected. The details of definitions of all graphical features and attribute information can be referred to “1:1,000 Basic Mapping Specifications (version 4.1)”. The general rules for data extractions include:

Feature Definitions:

Feature definitions provide the distinguish characteristic needed to differentiate between features. Although the difference between STREAM/RIVER and LAKE/POND is obvious, the distinction between them may not be obvious. Therefore, a quantitative measure is used to distinct STREAM and RIVER, for example, the width less than 0.5m is defined as STREAM.

Attributes and Attribute Values:

Definition of attributes and attribute values are generic. For example, the definition for Z_value is the vertical distance referenced to the Hong Kong Principle Datum.

Representation Rules:

The representation rules and definitions of all symbols showed on map face are described in legend of standard survey sheet.

Capture Conditions:

The data capture definitions are stated in “1:1,000 Basic Mapping Specifications (version 4.1)”. Capture conditions are generally independent for each feature. For example, the minimum area of building feature is 4 m^2.

Data Quality

At least five data quality elements relating to iB1000 shall be included in data quality report and metadata. The usage of data quality information is to allow users to have the freedom to evaluate the usefulness of the data, rather than applying quantitative threshold towards the dataset.

Lineage:

In the early 1950s, a set of map covered the New Territories area was produced at scale 1:1,200 which aimed to supplement the old Demarcation District sheets survey in the early 1900s….. (refer section 4.1.1)

Positional Accuracy:

iB1000 is produced using diverse data capture methods. Their positional accuracies are normally compiled to meet “SMO MAP ACCURCY STANDARDS” where 95 percent of well-defined feature are to be within 0.4 m of true mapped ground position. The data processing error, such as snapping, trimming and merging, are located within 0.01 m. The maximum positional adjustment for any nodes and associated line elements would be 0.005 m.

(Remarks – It is suggested developing a SMO MAP ACCURACY STANDARDS for i-Series product. This standard must define the positional accuracy of all spatial features (i.e. horizontal and vertical accuracy threshold), and provide data assessment procedure for quality evaluation during data capture, data updating and data conversion.)

Attribute Accuracy:

All attribute codes of iB1000 are agreed within 95 % to attribute code as described in data dictionary.

Logical Consistency:

Certain point-line-polygon relationship are collected or generated to satisfy topological requirements. Some of these requirements included: line begins and ends at nodes, line connects to each other at nodes, and lines do not extend through node, etc.

Completeness:

Completeness refers to commission and omission error. (refer section 1.2.5)

5.1.3 Data Quality Control

As discussed in Chapter in 4, SMO does not have a standard data quality assessment procedure to evaluate the quality of data obtained from external source. Figure 5.1 shows the conceptual design of data quality evaluation plan. For those external sources, a quality assessment procedure must conduct to determine whether the data is fulfil the requirements as stipulated in production specifications and data capture standards. Within these processes, the quality assessment procedure develop in section 3 can be used and the data standard proposed in section 5.2 can be adopted in quality assessment plan.

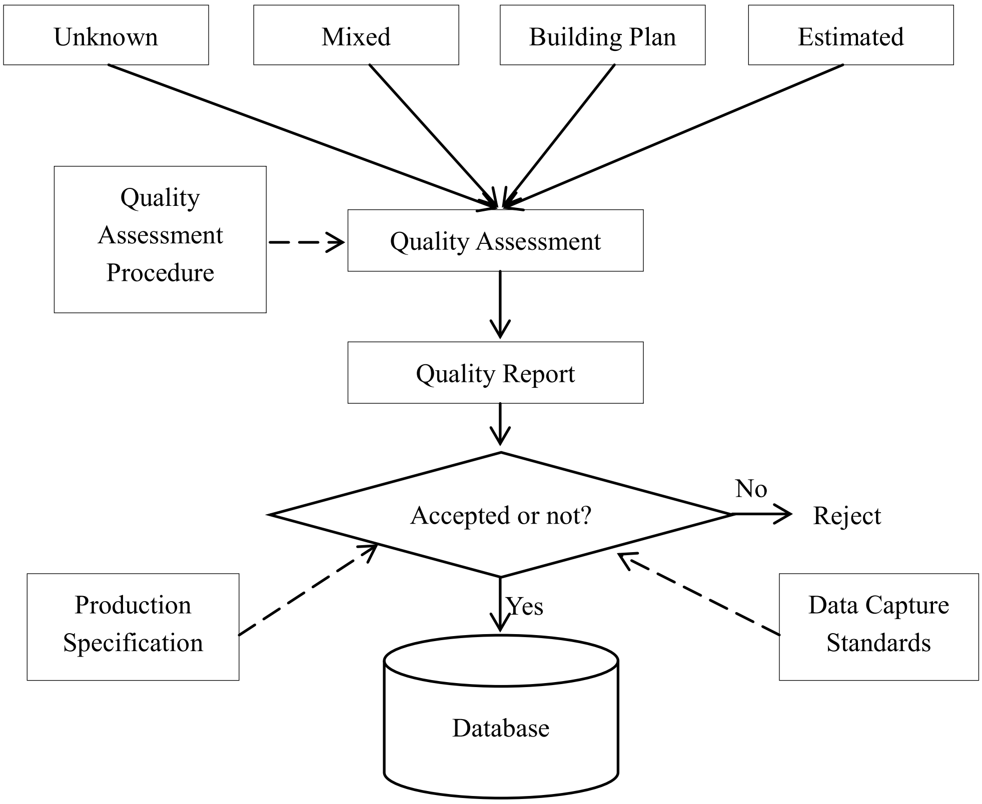


Figure 5.1 Conceptual design of data quality evaluation plan

5.1.4 Error Handling

SMO conducts QA/QC processes periodically to check the quality of products. In addition, SMO will perform data cleansing tasks for different situations, for examples, errors reported by data users, changes arose on product specifications, errors detected by periodical QA/QC processes and errors inhered in source materials, etc. There are various methods to cope with the errors, for examples, root cause analysis, correct current errors and prevent future data errors, etc. Gathering background information and clearly understanding the problems can help to find the root causes and prevent future errors.

5.1.5 Implement Control

A control process is a feedback looping through which the actual performance is measured comparing with standard. This can be done by assessment, inspection and review. SMO shall implement various controls to measure the quality of dataset, such as monthly quality check, users' feedback, and customer services, etc.

5.2 Applications of Quality Assessment Procedure in Data Quality Controlling System

As a central mapping agency in Hong Kong, SMO manages large amount of spatial data which acquired by internal survey teams and converted from external data sources. This is an effective and efficiency methods to maintain a large spatial dataset. This is also why many governments throughout the world have recognized the need to establishing spatial data infrastructure to maximize the data usage and avoid duplication effort in data capture. However, data obtained from diverse data sources may lead to problems in data integration and migration. For examples, Highway Department maintains a set vector-based RoadPolygon which facilitating road maintenance and construction work. This dataset may not fulfill the requirements for topographic representation that aims to manifest ground feature only. In addition, there may not have adequate quality information describing the quality elements such as data usage, positional accuracy and attribute accuracy, etc. Therefore, it is crucial for data providers to establish a framework to determine whether or not the obtained dataset be fit for use in their particular applications.

In order to have fruitful understanding of spatial data and make decision on the data acceptance, it is necessary to conduct quality check for the dataset. In this connection, data provider is recommended using the quality assessment procedures in different data capture methods which includes the field survey, photogrammetry survey and other data conversion processes, etc.

iB1000 is captured in accordance with the 1:1000 Basic Mapping Specifications (version 4.1) which aimed at maximizing consistency in the data content, structure and quality. The mapping specifications contain the definitions of spatial data based on users' requirements, for example, the selection criteria of what real world features are to be represented and how they are represented and positional accuracy requirements, etc. These definitions provide the basis for quality assessment and are considered as objectives in data quality assessment procedure.

This section proposes three applications of quality assessment procedure which is used in field survey, photogrammetry survey and data converted from external data sources.

Quality Assessment for Field Survey Data:

As mentioned in section 4.1.2, SMO uses total stations and GPS as field survey methods. The field surveys adopts Hong Kong 1980 (HK80) and the World Geodetic System (WGS84) as the horizontal control, and the Hong Kong Principle Datum (HKPD) as the vertical control. This ensures the absolute accuracy of control network.

Field survey also involves instrumental survey techniques which are distance measurements and angle measurements. For distance measurement, the errors can be classified into zero error, cyclic error and scale error. For angle measurement, the errors can be classified into eccentricity of centres, collimation in azimuth, transit axis error, circle graduation error, optical micrometer error and vertical circle index error.

Besides the errors in control network and survey equipment, the errors in radiation shall be taken into account, such as non-vertically of reflector pole.

Based on the above information, the quality assessment of field survey is preferred using deductive estimate methods which mentioned in section 2.2.6.

Quality Assessment for Photogrammetry Survey Data:

SMO has a photogrammetry section which capture aerial photos for various applications, such as digital orthophoto, change detection, and topographic map production, etc. For the purpose of topographic mapping, SMO uses scanned aerial photos with different scale and GSD, for example, scale 1:6,000 and GSD 7.5 cm.

Normally, accuracy of photogrammetry survey can be deduced by using the parameter of flying height, focal length, photo scale, scan resolution. But, it preferred using the methods indicated in NSSDA (1998) to test the accuracy. It suggests that the horizontal accuracy shall be tested by comparing the planimetric coordinates of well-defined points in the dataset with the coordinates of the same points form independent source of higher accuracy.

Quality Assessment for Data Converted from External Data Sources:

Currently, SMO converts data from external data sources, such as as-built building plans, engineering drawings and utility points captured by contractor, to create and update iB1000. Since some of these data sources may have unknown accuracy, a quality test shall be conducted to evaluate the conformance.

After data conversion, it is suggested using check points to test the positional accuracy and automatic program to test the attribute accuracy and logical accuracy. The positional accuracy is tested by comparing the coordinates of features within iB1000 to the coordinates of same points as determined by field survey. The attribute accuracy and logical accuracy is tested by conducting automatic program checks that developed by in-house staff.

CHAPTER 6

CONCLUSIONS AND FUTURE DEVELOPMENTS

Quality assessment and control of spatial data have long been a topic of study for many national mapping agencies and researchers. This study focuses on quality assessment of vector-based topographic data and quality controlling system for data providers to monitor the data quality throughout the entire data production process. This study analyzes the existing positional accuracy standards and positional accuracy assessment methods, and develops a quality assessment procedure for data providers and users to evaluate the quality of topographic data. Furthermore, this study reviews the Hong Kong 1:1,000 i-Series Digital Topo Map and conducts a case study by utilizing the proposed quality controlling system to manage the data production process. This chapter summarizes the findings of this study, and looks forward to the future developments of spatial data quality assessment and control.

6.1 Conclusions

The positional accuracy is one of the significant quality elements to describe the spatial data. The first part of this study presents an analysis of seven positional accuracy standards and several positional accuracy assessment methods, taking account of the sampling size, statistical formulation, testing methodologies, applicability, and reporting methodologies.

Theoretically, the greater the sample size the greater precision of the test, but the collection of sample data is expensive. Therefore, there is a need to balance these two factors. Among the seven positional accuracy standards, the recommended sample size is between 20 to 100. ASPRS (2014) suggested that the condition of selecting how many numbers of check points is depended on the testing area, for example, 40 horizontal check points for 1,500 km^2. In addition, the distribution of check points is depended on the shape of the testing area and the interest of the features, a random or equally distribution is recommended by most positional accuracy standards.

Regarding the statistical formulation and testing methodologies, most positional accuracy standards suggested that the testing for horizontal accuracy is conducted by comparing the coordinates (x and y) of well-defined points to the coordinates of the same points as determined by check survey of higher accuracy. Since elevation data rarely contains well-defined points, Triangulated Irregular Networks (TINs) of elevation data are interpolated at horizontal coordinates of vertical check points to determine the differences. In order to minimize the interpolation errors, the vertical check points are preferred locating at flat or uniformly-sloped terrain. The horizontal and vertical accuracy of spatial data are determined by statistical evaluation of random errors and systematic error and specified by root-mean-square error at either 90% or 95% confidence level. It is necessary to point out that ICSM (2009) is one of the standards suggested using single buffer overlap to test the horizontal accuracy of line features in dataset.

As the magnitude and distribution of errors of LiDAR-derived elevation data are closely related to the land cover types, ASPRS (2004) classified the land cover into open terrain, tall weeds, brush land, forested areas fully covered by trees and urban areas with dense man-made structure, and suggested the vertical accuracy in open terrain is tested to either 95% confidence level (normally distributed error) or 95th percentile (not necessary normally distributed) for different land cover types.

Although these positional accuracy standards have presented the fundamental issues to assess the positional accuracy of spatial data, it also encompassed some failures. The first one is that the horizontal accuracy assessment is still relied on testing of well-defined point features in dataset. Line segment is the most common geometry type which constitutes linear features and area features in a dataset, but there is lack of methods to assess and describe the positional accuracy of line segment. This study has reviewed several models representing the positional accuracy of line segment which shall be included in the new positional accuracy standards.

The second is that there is lack of sampling procedures and relevant examples presented in positional accuracy standards. Sample size is often the most critical aspect for quality assessment, because it links to the cost of test. The recommended minimum sample size is usually very low in positional standards, for example, 20. So, there is a need to create a sampling procedure which takes account of the shape, size, and terrain of the testing area.

The last is that most positional accuracy standards are focus on assessing vector, raster and several types of digital terrain model. There is seldom positional accuracy standards provide methods to evaluate new type of spatial data, such are 3D building model, 3D mesh model and street view videos. These products are widely used in most GIS and mobile applications, so there is a need to develop the corresponding spatial data standards.

ISO 19114 has provided a framework of procedures for assessing quality of spatial data, but there is lack of examples and methods to demonstrate the underlying principles. The second part of this study proposes a quality assessment procedure containing four steps, create data quality objectives, develop data quality assessment plan, determine data quality result, and report data quality result, which aims to incorporate new elements for quality assessment.

In the first step, the quality objective is created by identifying both internal and external requirements of the dataset. This information refers to how the dataset fulfills the requirements stated in product specifications and users' expectations. A dataset hierarchy, communication mechanism between data provider and user, and principles to create new data quality elements has been proposed to demonstrate the applications. The second step includes methods for identifying well-defined features in positional accuracy assessment, and introduces a sampling plan for evaluating attribute accuracy, completeness and logical consistency. The third step determines the counting-related data quality result and uncertainty-related data quality result of the quality assessment, and the last step give examples to report data quality by metadata and data quality report.

SMO has implemented a quality control system, but there is no data quality assessment plan and data capture standards. The last part of this study develops a quality controlling system which aims to strengthen the quality control system of i-Series TOPO MAP. The proposed quality controlling system encompasses five components which are data policy, data standards, data quality control, error handling and implement control. This system provides a standard quality assessment procedure which can be applied for evaluating the data acquired from field survey, photogrammetry survey and external sources, such as as-built building plans and engineering drawing. Furthermore, a data standard has been drafted which includes the spatial and non-spatial components. The new system allows SMO to have an overview of iB1000 and provide suggestions on how to improve the quality of the products and create a comprehensive data standard for users.

6.2 Future Developments

Positional accuracy assessment of line feature is a hot topic among researchers and many national mapping agencies. During the past 30 years, a numerous models have been developed to assess and describe the positional accuracy of line feature. However, these models are rarely included in contemporary positional accuracy standards and commercial GIS software, and most layman users may not familiar with their theories and real applications. A possible reason is that these models often have high level of complexity in terms of calculations and assumptions which make it difficult to implement in dairy GIS applications. In order to relieve these problems, it is suggested using a widely accepted method to assess and describe the accuracy of line feature in spatial data. The single buffer overlay is recommended for such purpose. Furthermore, it is suggested include this method in most positional accuracy standards and commercial GIS software for easy reference.

Development of spatial data infrastructure has been discussed in Hong Kong more than 10 years. A system named "Data Alignment Measures" which aims to provide a platform to share spatial data within government departments had been implemented in 2004. However, unclear data policy and lack of technical standards have minimized the usages of the system. In the forthcoming future, there is a need to review the system in different aspects, for examples, develop comprehensive product standards, avoid unnecessary duplication of data, facilitate diverse applications and enable data integration by users, etc. Data quality and clear quality information are significant issues in data sharing, so it is suggested developing a data quality transfer standard which includes data models, quality information, and other spatial and non-spatial information. This standard shall be agreed by all data owners and review periodically to reflect the changes in system requirements.

Challenges for the future from a data provider's perspective might focus on assessing the fitness for use. With increasing use of mobile mapping system, spatial data is widely used by layman users. It is necessary to inform users of spatial data quality in meaningful and understandable way. The feedbacks from data user are the best way for assessing the quality of data.

REFERENCE

1:1,000 Basic Mapping Specifications (version 4.1), 2015. Survey and Mapping Office, Lands Department, H.K.S.A.R. Unpublished material

ASPRS, 1990. ASPRS accuracy standards for large-scale maps. American Society of Photogrammetry and Remote Sensing, Photogrammetric Engineering and Remote Sensing. Vol 56, No. 7, pp 1068-1070

ASPRS, 2004. ASPRS Guidelines, Vertical Accuracy Reporting for LiDAR data . American Society of Photogrammetry and Remote Sensing. May 24, 2004

ASPRS and ASCE, 1994. Glossary of the Mapping Sciences. ASPRS, Bethesda Maryland and ASCE, New York

ASPRS, 2014. ASPRS Accuracy Standards for Digital Geospatial Data - Draft. American Society of Photogrammetry and Remote Sensing (Draft for Review), Photogrammetric Engineering & Remote Sensing

Aalder H.J.G.L., 2002. The Registration of Quality in a GIS. In Spatial Data Quality, Chapter Thirteen. London: Talyor & Francis , pp. 186-199

Ariza-Lopez F.J., Mozas Calvache A.T, 2012., Comparison of four line-based positional assessment methods by means of synthetic data, Geoinformatica,2012, 16:211-243

Chan K.K., 1998. Accuracy Report in the BMS. Unpublished Material

C. Victor Wu., 1994. Spatial Data Quality and Its Evaluation. Comput., Environ. And Urban System, Vol 18, No. 3, pp. 153-165

CEN/TC 287, 1998, PT, 05, Draft Quality Model for Geographical Information, Working Paper D3, January 1995

CDGM, 1998. Content Standard for Digital Geospatial Metadata. Metadata Ad Hoc Working Group, Federal Geographic Data Committee

Dassonvile L., Vauglin F, Jakobsson A. and Luzet C., 2002. Quality Management, Data Quality and Users, Metadata for Geographic Information, Spatial Data Quality. London: Talyor & Francis, pp. 202-215

Devillers R., Bedard Y., Fisher P., Stein A., Chrisman N., Shi W., 2010. Thirty Years of Research on Spatial Data Quality: Achievements, Failures and Opportunities, Transactions in GIS, 2010, 14(4): 387-400

District Survey Office Technical Manual (version 4.2), 2014. Survey and Mapping Office, Lands Department, H.K.S.A.R. Unpublished material

DLGS, 1999. Part 2 Specifications: Standards for Digital Line Graphs, [pdf] Department of the Interior U.S. Geological Survey National Mapping Division. Available at: http://nationalmap.gov/standards/pdf/2dlg0999.pdf [Accessed 3 Apr. 2016]

D. Rodolpehe, J. Robert, 2006. Fundamentals of Spatial data Quality, ISTE Ltd, USA

Dutton G., 1992. Handling positional error in spatial database. In proceedings of the 5th International Symposium on Spatial Data Handling, Columbia, SC: International Geographic Union, 460-469

E.P. Baltsavias, 1999. Airborne laser scanning: basic relations and formulas, ISPRS Journal of Photogrammetry & Remote Sensing 54 (1999) 199-214

FGCC, 1984. Standards and Specifications for Geodetic Control Networks. Federal Geodetic Committee, Sept

FGDC-GPAS, 1998. FGDC-STD-007, Geospatial Positional Accuracy Standard, Federal Geodetic Control Subcommittee, Federal Geographic Data Committee, Reston

Goodchild, M,F., Hunter, G.J., 1997. A simple positional accuracy measure for linear features, Int. J. Geograph. Inf. Sci., 11(3), 299-306

Greenwalt C.R. and Schultz M.E., 1962 and 1968. Principles of Error Theory and Cartographic Application. United States Air Force. Aeronautical Chart and Information Centre. ACIC Technical Report Number 96. St. Louis, Missouri. 60 pages plus appendices. This report is cited in the ASPRS standards as ACIC, 1962

Guptill S.C., Morrison J.L., 1995. The Elements of Spatial Data Quality, Oxford Elsevier

Harding J., 2006. Vector Data Quality: A Data Provider's Perspective. In Fundamentals of Spatial Data Quality, ISTE Ltd, Chapter 8, p141-159

Heo J., Woo J., Sang J.S., 2008. New line accuracy assessment methodology using nonlinear least-squares estimation. J Sur Eng 134(1):13-20

Hunter G.J., 1991. Processing Error in Spatial Database: The Unknown Quantity, Symposium on Spatial Database Accraucy, Melbourne, Australia

ICSM, 2009. Australian Map and Spatial Data Horizontal Accuracy Standard. Intergovernmental Committee on Surveying and Mapping, Australia

ISO 2859-1, 1999. Sampling procedures for inspection by attributes – Part 1: Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection. Geneva

ISO 3951-1, 2013. Sampling procedures for inspection by variables – Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL, British Standard

ISO 3951-5, 2006. Sampling procedures for inspection by variables – Part 4: sequential sampling plans indexed by acceptance quality limit (AQL) for inspection by variables (known standard deviation), British Standard

ISO 19113, 2002. Geographic information – quality principles. International Organization for Standardization, Geneve

ISO 19114, 2003. Geographic information – Quality evaluation procedures. International Organization for Standardization, Geneve

ISO 19115, 2003. Geographic information – Metadata. International Organization for Standardization, Geneve

ISO/TS 19138, 2006. Geographic Information – Quality measures, Geneve

ISO 8402, 1994. Quality management and quality assurance in design, development, production, installation and servicing, (www.iso.ch)

NMAS, 1947. National Map Accuracy Standards, Office of Management and Budget, Washington, D.C.

NSSDA, 1998. Geospatial Positional Accuracy Standard, part 3. National standard for spatial data accuracy, Federal Geographic Data Committee, Reston

NTMS, 2012. Geoscience Australia Topographic Data and Map Specifications for the National Topographic Database & NTMS Series 1:250,000 & 1:100,000 scale Topographic Map Products. Available at: http://www.ga.gov.au/mapspecs/topographic/v6/index.html [Accessed 3 Apr. 2016]

OSMM, 2016. OS MasterMap Topography Layer, User Guide. [pdf] Ordnance Survey.

Available at:
https://www.ordnancesurvey.co.uk/docs/user-guides/os-mastermap-topography-layer-user-guide.pdf [Accessed 3 Apr. 2016]

Shi, W., 1994. Modelling positional and thematic uncertainty in integration of GIS and remote sensing. ITC Publication, No. 22, Enschede

Shi, W., 1998. A generic statistical approach for modeling error of geometric features in GIS. Int J Geogr Inf Sci 12(2): 131-14

Shi, W., Liu W., 2000. A stochastic process-based model for the positional error of a line segment in GIS. Int J Geogr Inf Sci 14(1):51 – 66

Shi, W., Fisher P. (eds.), Goodchild M.F., 2002. Spatial data Quality. London: Talyor & Francis

Shi W. 2009. Principles of Modelling Uncertainties in Spatial Data and Spatial Analyses. Taylor & Francis Group

SDTS, 1998. American National Standards Institute, Information Technology – Spatial Data Transfer Standard (STDS) (ANSI-NCITS 320:1998): New York, New York

W.G. Cochran, 1977. Sampling Techniques (third edition), John Willey & Sons, New York