Truthing Production Data Capture

K. Bradley Paxton, Ph.D., Steven P. Spiwak, and Douglass Huang
ADI, LLC, 200 Canal View Boulevard., Rochester, NY 14623

Abstract
In order to really know how a production forms data capture system is doing, it has been customary to have keyers sample captured data fields and do "double key and verify" operations to determine the correct answers ("truth") of production data. In the system we call *Production Data Quality*, which will be used in the 2010 Census, we use software automation and good statistical design to reduce the human effort involved by as much as 40 times and obtain high quality "truth". Once the "truth" is known, the production data may be scored using whatever correctness criteria are appropriate for the application, for example, some type of a "soft match".

Key Words: Forms Processing, Data Capture, Truth, Data Quality, Independent Random Variables

1. Background

The objective of the *Production Data Quality* (PDQ) system is to quickly and cost-effectively determine the "truth" of handprinted fields captured from forms in a production environment, enabling the quality of the captured data to be determined.

The PDQ process consists of three main steps:

1. Form Completion, wherein the Provisional Truth is formed based on an independent Optical Character Recognition (OCR) engine with a human keying the rejects.
2. The Production Data is compared with the Provisional Truth on a hard match basis
3. Truth Scrubber, wherein discrepancies in Step 2 are resolved by human analysts

The bulk of the PDQ benefit in terms of determining the "Truth" of the real Decennial Response Integration System (DRIS) Production Data happens in Step 2. At that point, based on Census Dress Rehearsal experience, it turns out that about 92% of the data is sent to the "Truth" file, and about 8% goes into Truth Scrubber. We expect these numbers to improve for 2010 production.

These steps are shown in two block diagrams, on the following two pages:

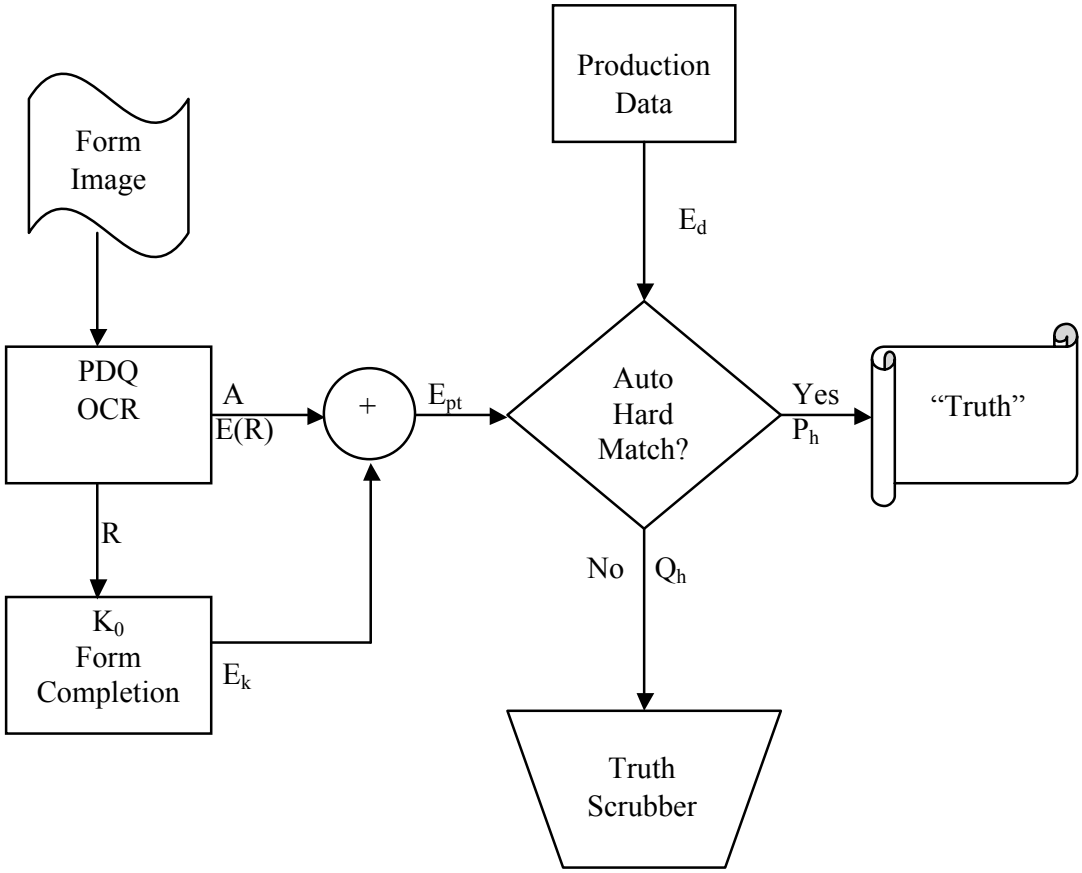

Figure 1: Form Completion and the First Match between Provisional Truth and Production Data

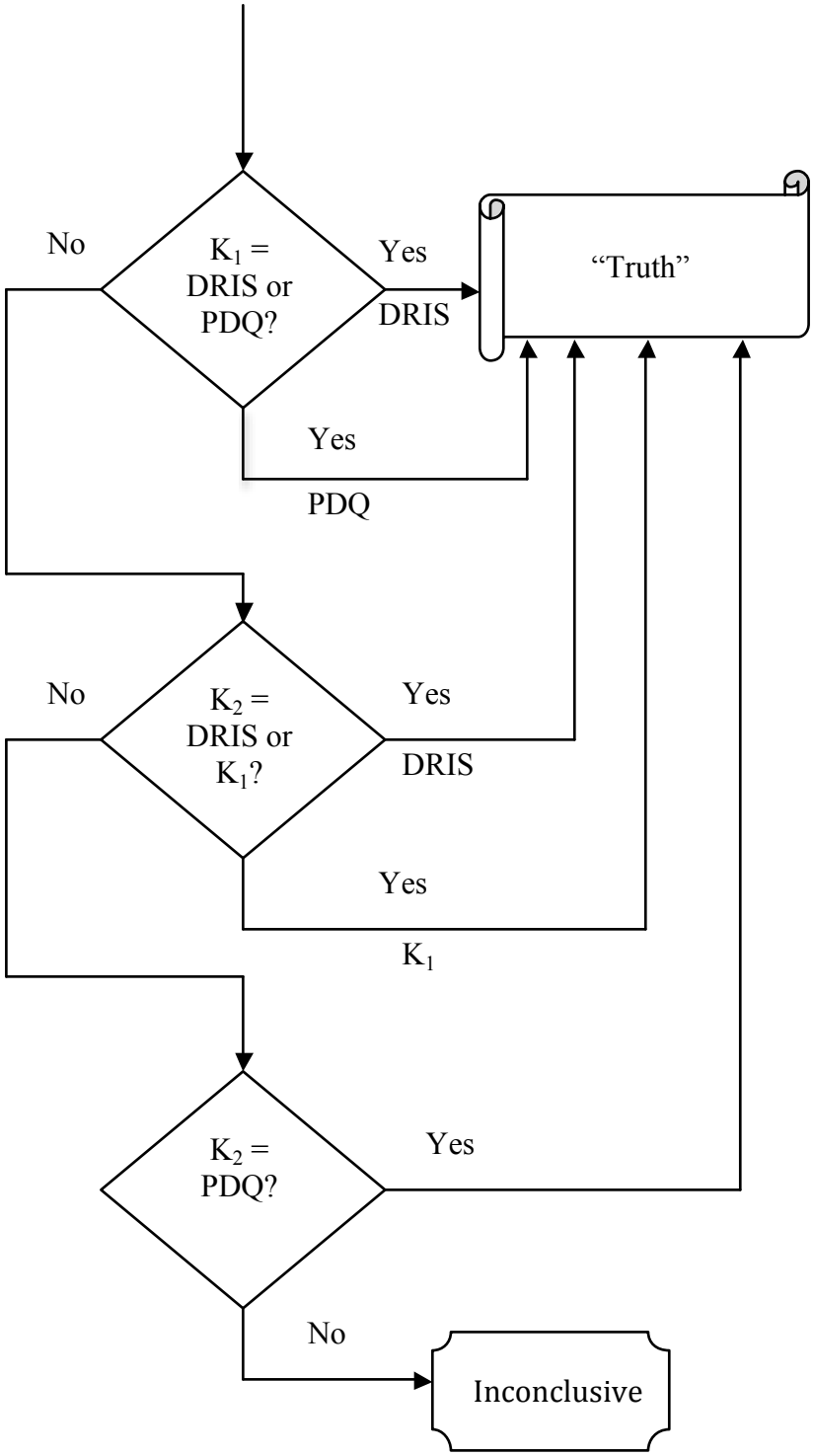

Figure 2: Truth Scrubber Using the Two Match Criteria

2. Analysis

We have developed some reasonably simple probability equations to describe how the (millions) of handprinted fields will move through PDQ during production. A key assumption in this analysis is that the various independently derived data sets, e.g., the Production Data, and the PDQ Provisional Data, are independent random variables (Ref. Parzen). Given this assumption, we can write, for two independent events A and B defined on the same probability space, the probability of both A and B occurring as P[AB] = P[A]P[B]. We have been using this assumption for over a decade in modeling data capture systems, but the advent of PDQ has finally made it possible to obtain data to verify this assumption regarding data capture in forms processing.

2.1 Summary of PDQ Equations

Referring to Fig. 1, we define the two basic field error rates as:

E_{pt} = Provisional Truth Error Rate

E_d = Production Data Error Rate

Using these definitions we can write the two basic equations describing the first match in PDQ as:

$$P_h = (1 - E_d)(1 - E_{pt}) \tag{1}$$

$$Q_h = E_d + E_{pt}(1 - E_d) \tag{2}$$

We define the error rates for the two Keyer/Analysts in Truth Scrubber as K_1 and K_2, respectively. Using the notation that the probability that the first keyer (K_1) agrees with production is P [K_1 = DRIS] referring to Fig.2, then we can also write explicitly all six of the additional probabilities pertaining to Fig. 2 (Truth Scrubber Equations) as shown below:

$$P[K_1 = DRIS] = (1 - E_1)(1 - E_d)E_{pt}$$

$$P[K_1 = PDQ] = (1 - E_1)(1 - E_{pt})E_d$$

$$P[K_2 = DRIS] = E_1E_{pt}(1 - E_2)(1 - E_d)$$

$$P[K_2 = K_1] = E_{pt}E_d(1 - E_2)(1 - E_1)$$

$$P[K_2 = PDQ] = E_1E_d(1 - E_2)(1 - E_{pt})$$

and finally, the probability of an inconclusive outcome is

$$P[I] = E_1E_2(E_d + E_{pt}) + E_dE_{pt}(E_1 + E_2) - 3E_1E_2E_dE_{pt}.$$

If you wish to do some messy algebra, you can convince yourself that $Q_h = 1 - P_h$ and that the six Truth Scrubber Equations shown above actually sum to Q_h. (It's not pretty, but it's correct).

Looking closer at the above equations, we see that the probability that both PDQ and DRIS being correct (P_h) is close to one; in mathematical terms this is often written as the expression $P_h = O(1)$, usually read "P_h is of order one". Following this analysis further, we see that the fields going to Truth Scrubber are $Q_h = O(E)$, that is, about the size of one of the (small) error rates. The first step in Truth Scrubber involves the first Keyer/Analyst K_1, and we see that both positive outcomes are close to an error rate, that is $P[K_1 = DRIS] = O(E)$ and $P[K_1 = PDQ] = O(E)$. This means that the first Keyer/Analyst "gets most of it" right away, and there is little work left to do. The probability that the second Keyer/Analyst contributes something useful to the "Truth" is of order E^2, because $P[K_2 = DRIS] = O(E^2)$, $P[K_2 = K_1] = O(E^2)$, and $P[K_2 = PDQ] = O(E^2)$. Finally, we get to the end of the Truth Scrubber Process, and what is not considered "Truth" at that point is essentially unknowable by PDQ: we call it "Inconclusive". The probability that a field gets to the end of the process as "Inconclusive" is of order E^3, and indeed we see that $P[I] = O(E^3)$.

2.2 Data Mined from PDQ (Alphabetic Write-In Fields)

Workflow data was extracted from PDQ in early May for a group of Census Dress Rehearsal "Short" (DX-1) forms consisting of 333,262 alphabetic write-in fields. This is about half of all the alpha fields processed in Dress Rehearsal. When the Provisional Truth was hard matched with the Production Data, it was found that 306,653 of them matched (92.02%), and were sent directly to the "Truth" file and 26,609 did not hard match (7.98%), and were sent to Truth Scrubber (see Fig. 1).

Further, it was found that the number of fields in error in the Production Data was 7,376 and the number of incorrect fields in the Provisional Truth was 19,718. We also found that 51 fields were Inconclusive at the end of the PDQ process, so the proper denominator for computing error rates is 333,262 – 51 = 333,211. (This is because we do not use inconclusive fields when computing error rates).

The error rates derived from the mined data are then:

$$E_{pt} = 19{,}718 / 333{,}211 = 0.0591757$$

and

$$E_d = 7{,}376 / 333{,}211 = 0.0221361.$$

Using Equations (1) and (2) and the above error rates estimated from mined data, we can compute the two probabilities of matching or not matching as:

$$P_h = (1 - 0.0221361)(1 - 0.0591757) = 0.9199981 = 92.00\%$$

$$Q_h = 0.0221361 + 0.0591757(1 - 0.0221361) = 0.0800018 = 8.00\%$$

So, the theory as expressed by Equations 1 & 2 agree with the actual data to within 0.02%. This strong agreement suggests that the assumption of the Provisional Truth and the Production Data being independent random variables is very good. We somewhat expected this because both of these data files are produced largely by computer automation with modest human assistance.

Digging deeper, of the 26,609 fields that went to Truth Scrubber, the mined data gives 18,905 fields for Keyer 1 agreeing with DRIS, 6,395 fields for Keyer 1 agreeing with PDQ, 136 fields for Keyer 2 agreeing with DRIS, 1,093 fields for Keyer 2 agreeing with Keyer 1, 29 fields for Keyer 2 agreeing with PDQ, and finally, 51 fields left as inconclusive. We also did an independent estimate of keyer error rate in Truth Scrubber, and obtained 2.0935% (hard match).

Using the above data, we can then make a table of theory vs. data as shown below in Table 1:

Table 1: Comparison of PDQ Theory with Data

Probability	Theory	Data	Data-Theory
P_h	92.00%	92.02%	0.02%
$P[K_1=DRIS]$	5.67%	5.67%	0.01%
$P[K_1=PDQ]$	2.04%	1.92%	-0.12%
$P[K_2=DRIS]$	0.12%	0.04%	-0.08%
$P[K_2=K_1]$	0.13%	0.33%	0.20%
$P[K_2=PDQ]$	0.04%	0.01%	-0.03%
P[I]	0.00%	0.02%	0.01%
Total	100.00%	100.00%	0.00%

Table 1 shows extraordinary agreement between the math model and the actual mined data from PDQ. This table gives the results for all six pathways to the "Truth" in Figs. 1 & 2, as well as the final path to the residual inconclusive fields. The high (as designed) productivity of PDQ arises from the fact that 92% of the data is "swept off the table" by the first auto hard match step, and an additional 7.6% of the data is handled by Keyer 1, so that only 0.4% remains to be handled by Keyer 2.

It is a somewhat fine point, but the largest error rate in a particular path between theory and data is 0.2%, and occurs for the case where Keyer 1 agrees with Keyer 2. In particular, we see that the data shows 0.2% more agreement between the two keyers than predicted by the model. We believe this may be because this is the only path where we are looking for <u>two</u> humans to agree (as opposed to a human and a computer)! There were some subtle keying "rules" that were found to be not uniformly followed during the 2008 Dress Rehearsal, from which this data was taken. For example, in the case where "N/A" is written in a write-in field, the keyer is instructed to leave it blank. So if two keyers both disobeyed the same keying rule, and typed "N/A", they could agree in a case where they were both incorrect.

We expect, based on recent engineering efforts, that this problem will much less pronounced in upcoming pre-production testing.

3. Conclusions

We have shown the assumption of independent random variables for the various independently derived data sets in PDQ is a very good assumption, based on the strong agreement of data with the probability math model.

We have indicated how the PDQ system is able to be a fast and cost-effective way to precisely determine production data capture quality.

We expect to see some improvement in this system during the final tests leading to production in 2010, however, it is already extremely good based on our Dress Rehearsal experience.

Acknowledgements

The authors appreciate the hard work of the ADI team in developing (and continuing to develop) PDQ. We wish to thank Fred Highland of Lockheed Martin Corporation for organizing the panel "Statistical Methods in Census Data Capture", and for suggesting this paper. We also appreciate many constructive technical data capture quality conversations with Dick Taylor of Evolver, Incorporated. Finally, we wish to thank Alan Berlinger of the U.S. Census Bureau for his diligent support of the many complex aspects of data capture quality.

References

Parzen, Emanuel, *Modern Probability Theory and Its Applications*, Wiley & Sons, New York, 1960.