Pitfalls In Studying "Big Data" From Operational Scenarios

Estefan Ortiz and Kevin W. Bowyer
Department of Computer Science and Engineering, University of Notre Dame

Abstract

Analyzing a larger dataset is sometimes assumed, in itself, to give a greater degree of validity to the results of a study. In biometrics, analyzing an "operational" dataset is also sometimes assumed, in itself, to give a greater degree of validity. And so studying a large, operational biometric dataset may seem to guarantee valid results. However, a number of basic questions should be asked of any "found" big data, in order to avoid pitfalls of the data not being suitable for the desired analysis. We explore such issues using a large operational iris recognition dataset from the Canada Border Services Agency's NEXUS program, similar to the dataset analyzed in the NIST IREX VI report.

1. Introduction

NEXUS is a highly successful Canada Border Services Agency (CBSA) program for expedited crossing of trusted travelers at the US-Canada border [1]. The NEXUS program uses iris biometrics to recognize identity. In this paper, we analyze a version of the NEXUS iris recognition dataset that is a superset of that analyzed in IREX VI [2]. Our version includes additional metadata that gives deeper insight into the origins of the iris match scores

The IREX VI report is titled "Temporal Stability of Iris Recognition Accuracy" [2]. It reports conclusions that are at odds with those of other studies on iris template ageing [3,4,5]. Those other studies analyze smaller datasets acquired for research purposes. IREX VI suggests that analyzing a larger, operational dataset contributes to the validity of its conclusions [2]; e.g.,*"Using two large operational datasets, we find no evidence of a widespread iris ageing effect. Specifically, the population statistics (mean and variance) are constant over periods of up to nine years", "Our best estimate of iris recognition ageing is derived from a 7876 person subset of an operational registered traveler deployment …",* and *"In conclusion, we assert that operational logs of successful recognition attempts are an invaluable resource, not least because of their large size".*

Our analysis shows how big data, and especially big data from an operational scenario, can contain subtle and unanticipated complexities. We point out how these complexities can complicate analysis of such datasets to answer research questions. In particular, complexities that arise in a large, operational dataset can make it difficult to obtain a meaningful answer to an apparently simple research question such as iris template aging.

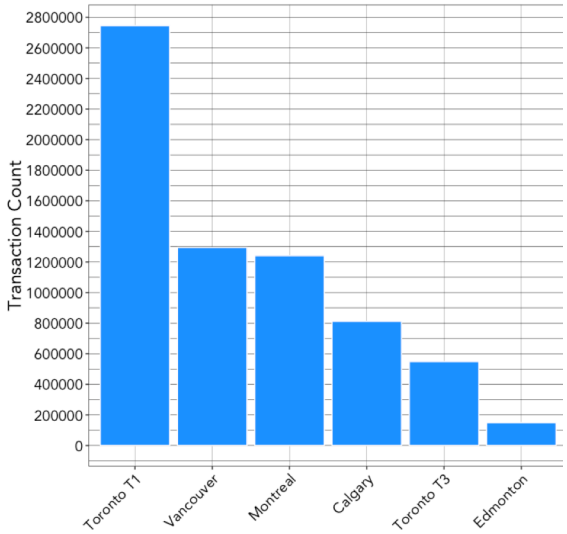

Figure 1. Border-crossing transactions by airport location. Five locations account for most of the data; lowest-traffic locations not shown on histogram.

2. One Data Stream or Many?

One potential pitfall associated with "big data" is that the bigger that the dataset, the greater the chances that it was collected at multiple points, at multiple times, or in multiple ways. This concern is especially important when the dataset is "found" data. By "found", we mean data created as a result of an operational scenario run with the goal of an efficient and user-friendly application, rather than a dataset whose collection was conceived and executed to support research. Therefore, a fundamental question to ask is – *Can the "big data" dataset be appropriately analyzed as one homogenous dataset?*

Table I of [2] describes its OPS-XING dataset as collected in 2003-2012, at airport border crossings, involving 521,474 eyes of 350,566 people, and containing 5,710,434 (assumed) genuine match scores. It does not mention the number of airports or kiosks involved. NEXUS is in fact a large, comprehensive program serving

Canada-US travelers at 11 locations: Toronto terminals T1 and T3, Ottawa, Vancouver, Montreal, Calgary, Edmonton, Halifax, Winnipeg, Fort Erie and Billy Bishop Toronto City Airport. As shown in Figure 1, Toronto T1 is the highest-traffic location, and five of the locations account for a large fraction of the data. Our analysis focuses primarily on data from these five locations.

At most locations, there are multiple iris kiosks. A traveler initiates a border-crossing transaction without making an identity claim. The iris image is acquired and the resulting iris code compared to those of enrolled travelers until an acceptable match is found. Once an acceptable match is found, the search ends and comparisons are not made to any remaining enrollment records. This "token-less" and "1-to-first" recognition process emphasizes convenience and speed for the traveler. If an acceptable match is found, meta-data recorded with the match score includes location, crossing direction (into CA or into US), kiosk number and other data. If no acceptable match is found, no data is recorded.

There are kiosks for traffic into CA and for into US, in different places in the airport. Figure 2 displays the mean normalized Hamming distance (nHD) by month for the five highest-traffic locations, collectively, broken out by into CA / into US. (The normalization step appears to be that proposed by Daugman [6] to adjust match scores for different numbers of bits participating in the iris match.) One surprise in Figure 2 is that *traffic into CA has on average a better iris match score than traffic into US!*

For a given month, the difference in mean nHD may be 0.005 or larger for travelers coming into CA versus into US. The annual rate of change stated as the "best estimate" of iris aging in IREX VI is 8×10^{-7} [2]. Thus the difference in sampling data from the CA stream versus the US stream, something not accounted for in the IREX VI analysis, is about four orders of magnitude greater than the effect that IREX VI claims to measure.

Another surprise in the data in Figure 2 is that *there is a seasonal pattern of highs and lows in the mean nHD!* The US data shows a low nHD in September of each year from 2009 to 2013, and a high nHD in December in four of those five years. The CA data shows a low mean HD in September or October, and a high typically in March. The difference in the seasonal high and low can be as large as 0.007 in HD; compare September to December of 2009 in the US data. The seasonal pattern in the data introduces an additional element of complexity for the analysis of any trend over time.

As shown in Figure 3, *there are also clear differences in mean nHD between airport locations!* For the CA locations, Vancouver generally has the highest mean nHD and Calgary the lowest. The difference between them is on the order of 0.01. For the US locations, Calgary and Vancouver generally have the highest mean nHD, and Montreal is generally the lowest.

Note that, not only are the data streams from the individual airport locations different from each other, they also differ in some respects from the aggregate data

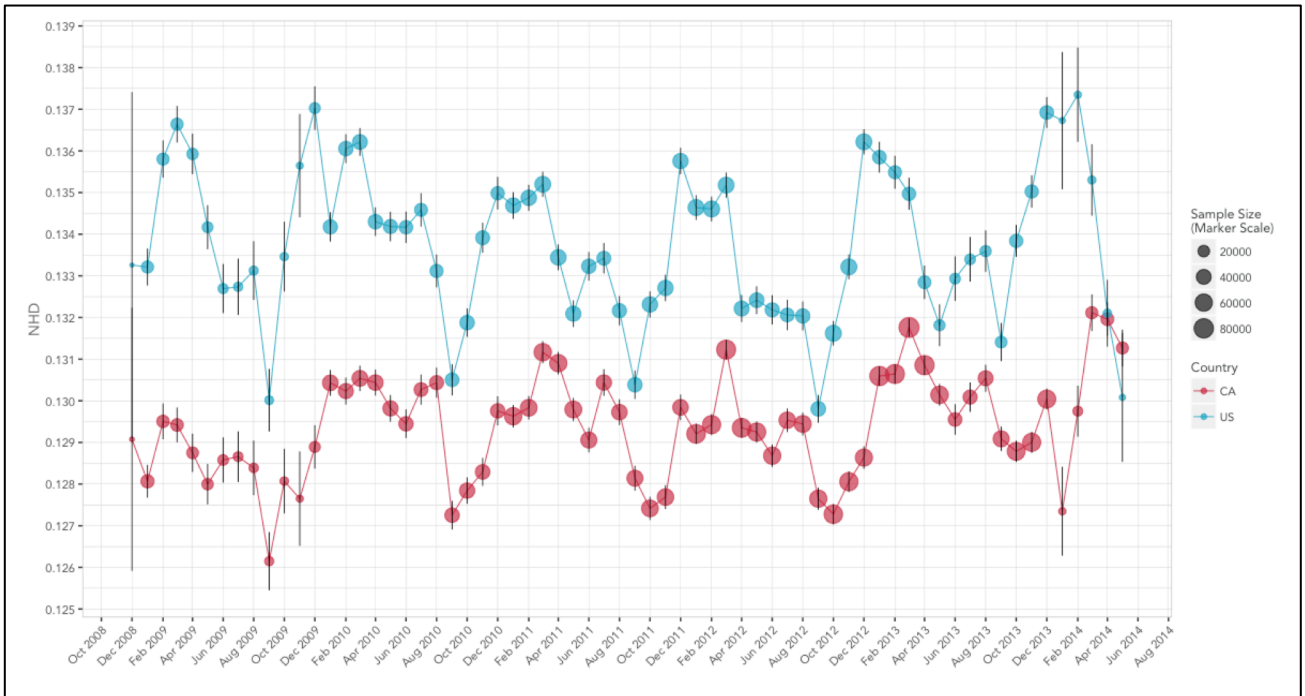

Figure 2. Mean nHD for left iris by month, five highest-traffic locations together, by US / CA. Circles indicate number of HD scores, bars indicate standard error of the estimate (SEE); note difference in US and CA, and seasonal high/low pattern.

stream. For example, the same clear pattern of seasonal min and max nHD in the aggregate US data in Figure 2 is not consistently repeated in all of the US individual location data streams in Figure 3.

The data streams from kiosks at a given airport location also show differences. Figure 4 illustrates this for three US kiosks at Toronto T1. The data streams for kiosks OK64 and OK65 are roughly similar, but the stream for OK66 represents a generally lower mean nHD than for either of the other two kiosks. The difference in mean nHD between two kiosks at the same location in a given month is as high as 0.007.

It seems clear that this big, operational dataset is a collection of individual datasets that have some significant differences. This of course raises the question of how to best take these differences into account in a study of, for example, iris template aging. The existence of such differences also raises a more important question: *What can we learn about the design at some airport locations in order to design more consistent and more accurate operational scenarios?*

3. Seasonal Variation In the Data

Another factor in "big data" becoming big is that it may be acquired continuously over time. If this is the case, it is important to ask if there is change in the data stream over time, or time-based patterns of variation in the data. We have seen that the dataset considered here exhibits a degree of seasonal variation.

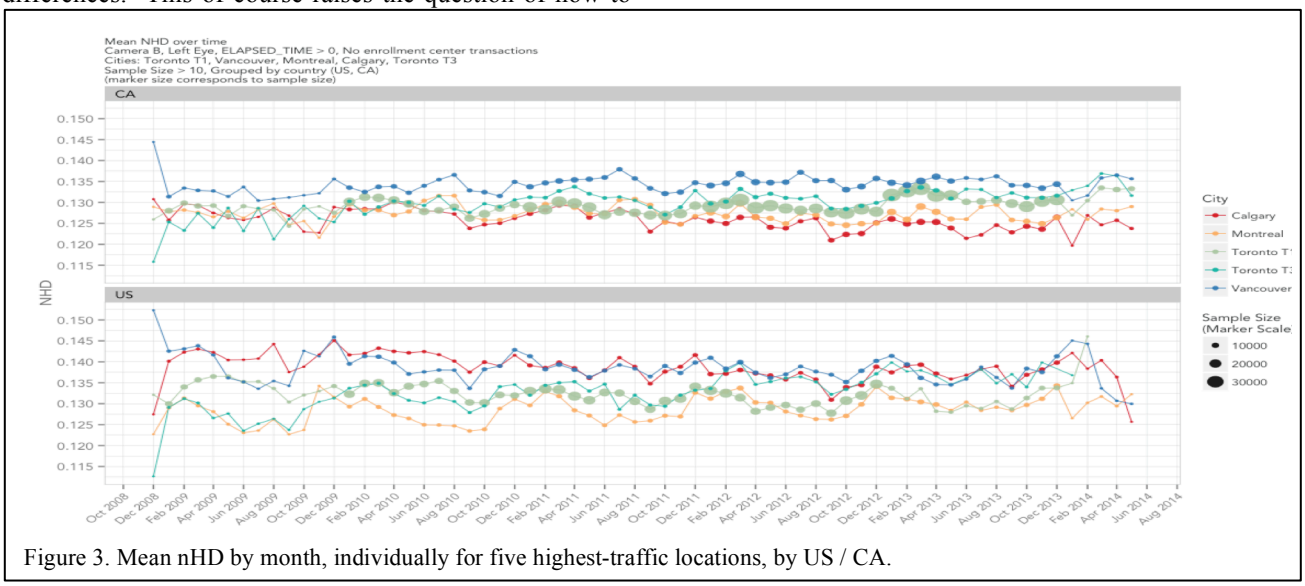

Figure 3. Mean nHD by month, individually for five highest-traffic locations, by US / CA.

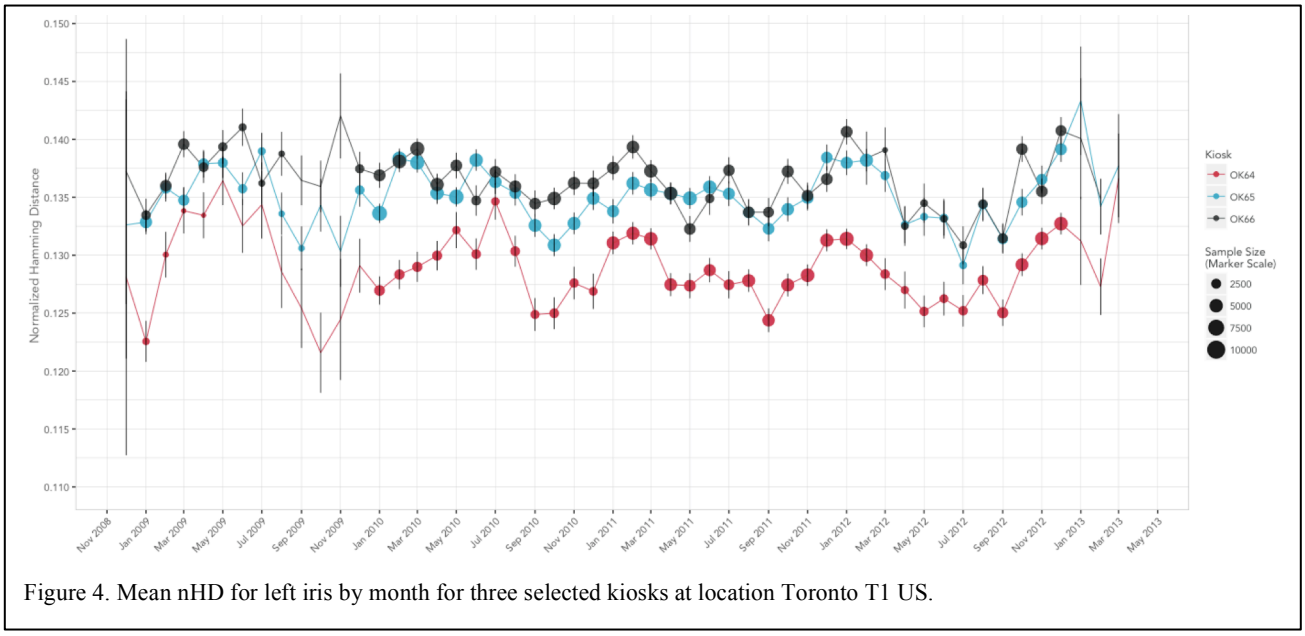

Figure 4. Mean nHD for left iris by month for three selected kiosks at location Toronto T1 US.

The seasonal variation in nHD may seem puzzling. However, the data contains clues to possible causes. Figure 5 shows the mean pupil dilation ratio of the probe iris image, broken out by CA / US. (Probe dilation ratio is an item of metadata recorded with each match score.) Note that the values for US transactions vary over a wider range than those for CA transactions. Note also that the mean pupil dilation for US transactions is at a low each year in July and at a high each year in December. Consider that in this location the mean number of hours of daylight is at its low in December and its high in July [7]. Low mean dilation occurs when seasonal daylight is high, and high mean dilation occurs when seasonal daylight is low. This suggests that at least some airport locations have kiosks placed where the seasonal level of daylight is affecting the pupil dilation.

Figure 6 shows the data in Figure 5 broken out by airport location. The pattern of pupil dilation clearly differs between airports. In the US data, Toronto T1 and Vancouver appear to have kiosks in locations where natural light plays a large role. Calgary, Montreal and Toronto T3 appear to have kiosks where natural light plays a lesser role, and Calgary appears to have the lowest level of ambient indoor lighting (causing highest dilation).

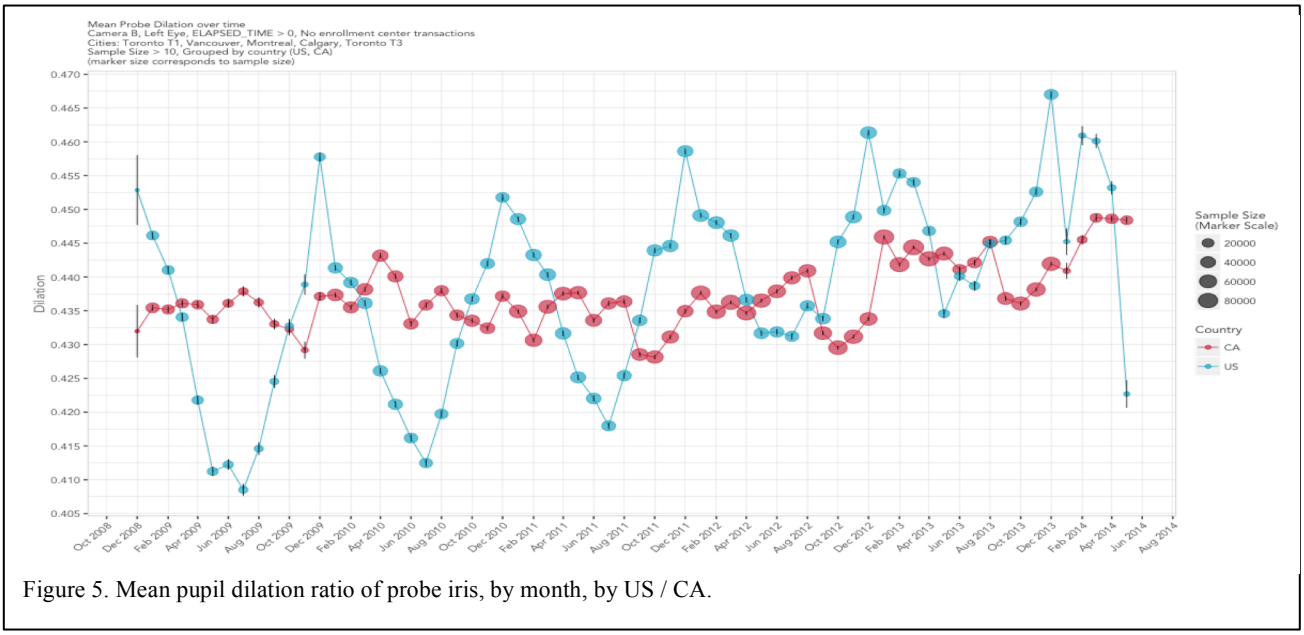

Figure 5. Mean pupil dilation ratio of probe iris, by month, by US / CA.

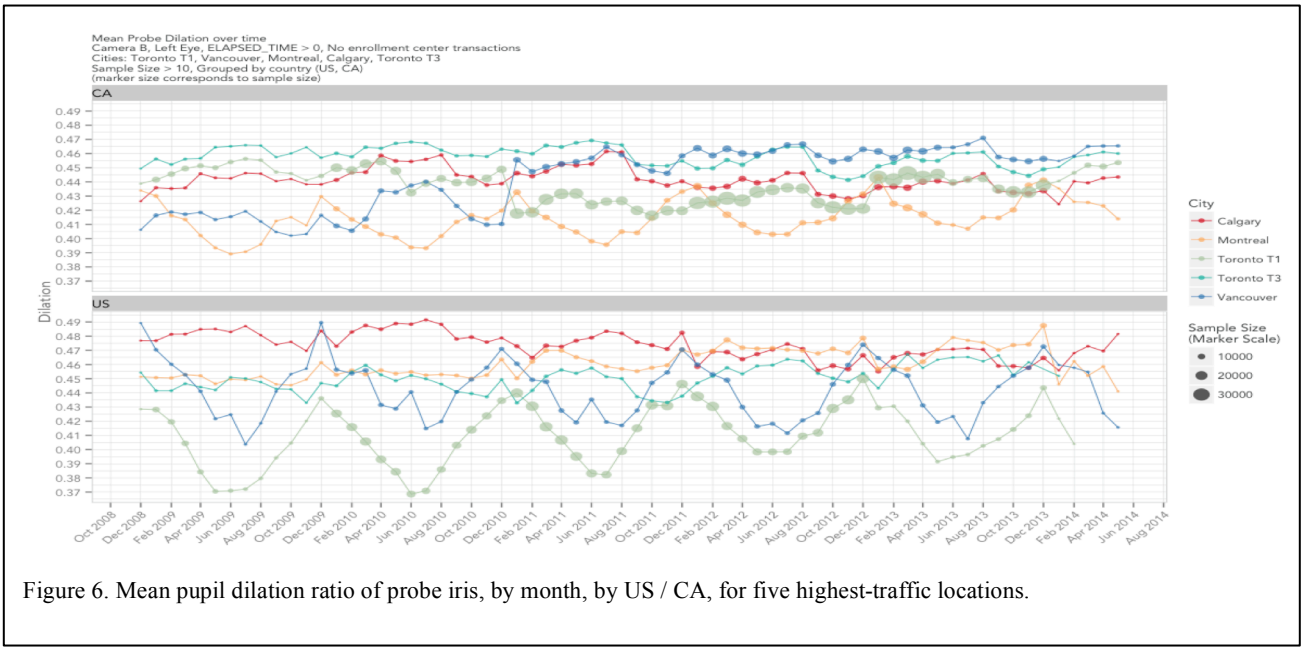

Figure 6. Mean pupil dilation ratio of probe iris, by month, by US / CA, for five highest-traffic locations.

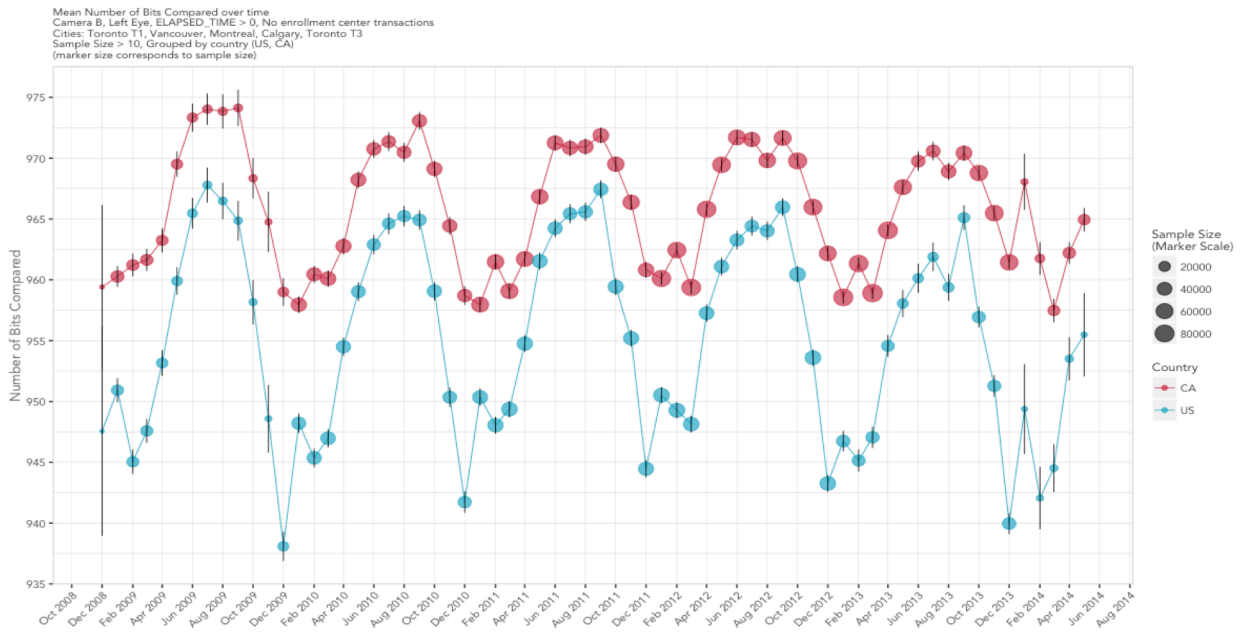

Figure 7. Mean number of bits in iris code match, by US / CA.

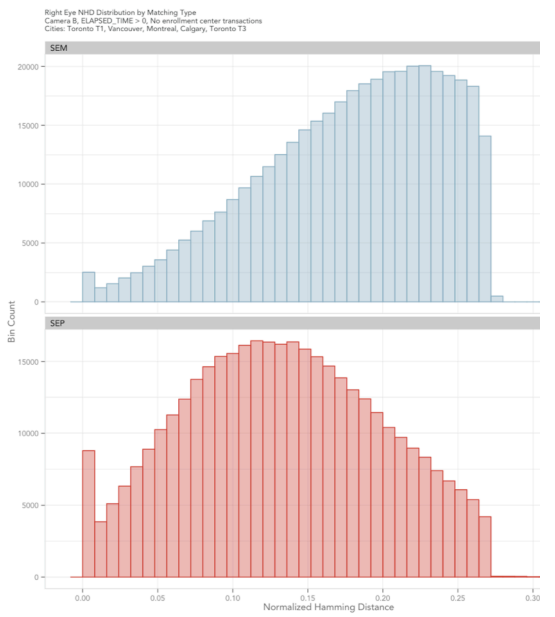

Figure 8. Distribution of assumed genuine right iris match scores for SEM (upper) and SEP (lower) acquisition mode. Left iris distributions (not shown) are essentially the same as the SEP (lower) right iris distribution. See Section 4.

Similar conclusions can be inferred from the CA data in Figure 6. Montreal appears to have the kiosks in an area influenced by natural lighting, whereas Toronto T3 seems to have used an indoor location with lower ambient lighting level. Looking closely at the Vancouver data, it appears that the ambient lighting level of the kiosk location was changed after December 2010, moving to a more controlled, lower-lighting-level scenario.

Figure 7 shows the seasonal variation in the number of bits used in a comparison of iris codes. This number is how many of the bits in the enrolled iris code and in the probe iris code were unmasked in both codes. Note the inverse relation in the number of bits and the level of pupil dilation. When dilation is highest, in December, the number of bits used in an iris code match is at its lowest; when dilation is lower, the number of bits is higher.

This suggests the possibility that pupil dilation in the probe image, the number of bits in the iris code match, and nHD can have some complicated inter-relationship that is affected by lighting at the probe kiosk. Lower light at the kiosk where the probe image is acquired causes greater dilation. And the difference between the probe and enrollment dilation, as well as high dilation in general, are known to cause an increase in the mean HD [8].

4. Is Data Collection Consistent Over Time?

The assumption in IREX VI is that the right iris scores were collected following the same conditional protocol over the lifetime of the operational scenario [1] – "*The eye itself, left or right, is influential: right eyes give higher HD values … This occurs because the right eye is used only if the left eye failed or was not acquired. The number of left eye events in the OPS-XING database is 4,920,638 vs. 725,300 for the right eye.*"

In actuality, while the left iris scores were collected consistently over time, the right iris scores were collected by two different protocols in different periods of time. (This is the reason that the analyses in this paper use only left iris data, rather than left and right.) In the "SEM"

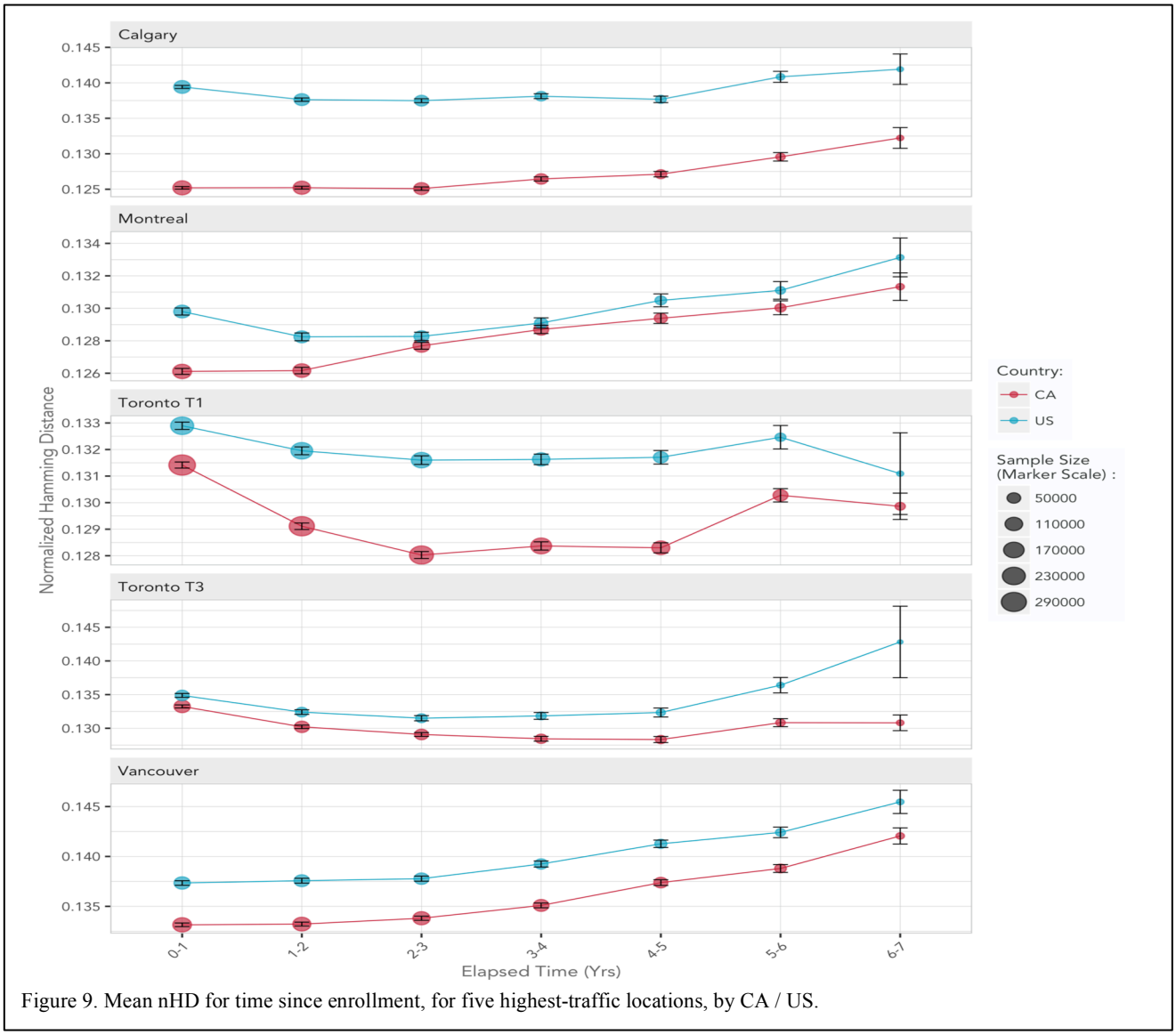

Figure 9. Mean nHD for time since enrollment, for five highest-traffic locations, by CA / US.

protocol described above, the right iris match score is generated only if there is not a successful match with the left iris. This protocol results in the right iris genuine distribution being worse than that for the left iris, as illustrated in Figure 8. In the other protocol, referred to as "SEP", the left and right iris match scores are both generated independent of the outcome of matching the left iris. The SEP protocol results in the right iris genuine distribution being the same as the left iris genuine distribution. The collection protocol, SEM or SEP, is an element of recorded meta-data. This property of the CBSA dataset was also noted in [9]. (For insight into the idea that the conditionally acquired score has a worse distribution, see [11].)

The important question illustrated here is – *Does the collection protocol vary over time in any known way*?

5. Discussion and Conclusions

One factor contributing to the bigness of "big data" may be that it is acquired in multiple locations. When data is acquired in multiple locations, there is the chance for variations in the data stream based on factors specific to the different locations.

Another factor contributing to the bigness of "big data" may be that it is acquired continuously over a period of time. When data is acquired over a long period of time, there is the possibility for seasonal variation in the data, as well as the possibility of change in locations, protocols or other factors in the data collection.

A factor that is especially relevant for operational data is that it is acquired with attention to specific goals other than research. The conditions of acquisition may change

to provide better service to the customer, to better accommodate the goals of the sponsoring agency, to deal with growth in the application, or for other reasons.

The general pitfall in each instance is in assuming that the big, operational data is simpler than it is in reality – assuming that there are no important differences across data collection locations, that there are no important differences of a time-varying or seasonal nature, or assuming that the acquisition protocol remained constant. Every such assumption needs to be examined to verify that it is true to the extent that it does not affect the analysis needed for the research question. This principle can be seen as an extension of the guidelines given by [10] about use of datasets in biometrics research.

Research datasets are typically smaller than operational datasets, and do not exhibit the rich complexity of operational datasets. This may make it hard to answer some questions. But it also may make it easier to employ analysis methods that are appropriate to the dataset. Operational datasets are typically (much) larger, and by definition have the full complexity of the real application. But that complexity may make it exceedingly difficult to find appropriate methods of analysis. And some operational datasets may simply have not recorded sufficient meta-data to even understand the complexity of the data. We are fortunate in this particular instance that the CBSA dataset was acquired with rich meta-data, allowing us a glimpse of the complexity involved in the data. The essential message is not that research datasets are better, or that operational datasets are better, but that methods of analysis need to be appropriately matched to properties of the dataset, in the context of the question to be studied.

One major conclusion from our analysis is that lighting is an important design element of an operational iris recognition scenario. The use of natural light in the acquisition area appears to be at the root of variations in the recorded nHD in this scenario. Placing kiosks where there is a large element of natural lighting may be motivated by considerations of having an acquisition environment that is user-friendly in aesthetic terms. However, for the technical goal of minimum variation in nHD, consistent low dilation at both enrollment and recognition is desirable, and so a significant component of natural light may not be good.

What, if anything, can be said about iris template aging based on this dataset? Figure 9 shows different patterns of change in nHD with time since enrollment at the level of the five highest-traffic locations. The first point on the plots is the mean nHD recorded within the first year after enrollment, the second point is the mean nHD recorded in the second year after enrollment, and so on. In some cases the pattern might be interpreted as strong template aging; for example, in Vancouver the change in mean nHD over seven years is about 0.008. In other cases, the pattern might be interpreted as an early improvement in nHDs, possibly due to users learning how to use the system, followed by a template aging effect setting in; for example, Toronto T1 or T3. However, given the number and variety of uncontrolled factors involved in the data collection, it would be speculative at this point to make any conclusion about "the" cause of any time-varying trend in the match scores.

In future research, we hope to develop a better understanding of the seasonal and location-dependent fluctuations in the iris match scores.

Acknowledgements

The authors express appreciation of the CBSA for their experience in running an exemplary biometric-enabled border-crossing program, and for their indulgence in allowing researchers access to their dataset.

References

[1] Canada Border Services Agency, http://www.cbsa-asfc.gc.ca/prog/nexus/air-aerien-eng.html

[2] P. Grother, J. R. Matey, E. Tabassi, G. W. Quinn and M. Chumakov, IREX VI: Temporal Stability of Iris Recognition Accuracy, NIST Interagency Report 7948, July 24, 2013.

[3] S. P. Fenker, E. Ortiz and K. W. Bowyer, Template Aging Phenomenon In Iris Recognition, *IEEE Access* 1, 266-274, May 16, 2013

[4] A. Czajka, Influence of iris template ageing on recognition reliability, *Communications in Computer and Information Science*, Volume 452, 2014, pp. 284-299.

[5] Sazonova, N., Hua, F., Liu, X., et al.: 'A study on quality-adjusted impact of time lapse on iris recognition'. Proc. of SPIE #8371B: Biometric Technology for Human Identification, 23–27 April 2012.

[6] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1167–1175, October 2007

[7] http://www.toronto.climatemps.com/sunlight.php

[8] K. Hollingsworth, K. W. Bowyer and P. J. Flynn. Pupil dilation degrades iris biometric performance, *Computer Vision and Image Understanding* 113 (1), January 2009, 150-157.

[9] E. Ortiz and K. W. Bowyer, Exploratory Analysis of an Operational Iris Recognition Dataset from a CBSA Border-Crossing Application, *CVPR Biometrics Workshop 2015*, Boston.

[10] A. Jain, B. Klare and A. Ross, Guidelines for best practices in biometrics research, *8th IAPR Int'l Conf. on Biometrics, 2015*.

[11] A. Czajka and K.W. Bowyer, Statistical evaluation of up-to-three-attempt iris recognition, *IEEE 7th Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS 2015)*.