IREX VI: Mixed-effects Longitudinal Models for Iris Aging: Response to Bowyer and Ortiz

Patrick Grother, James R. Matey and George W. Quinn
National Institute of Standards and Technology, Gaithersburg, MD, USA

***Abstract*—Bowyer and Ortiz, in their paper "A Critical Examination of the IREX VI Results", make seven criticisms of our application of linear mixed-effects models to longitudinally collected iris recognition Hamming distances. We reject these as either irrelevant, misinterpretations, or qualitatively correct but quantitatively irrelevant.**

***Index Terms*—biometric ageing; longitudinal analysis; iris recognition.**

I. Introduction

The Iris Exchange (IREX) program was initiated at NIST in 2008 to expand iris recognition capabilities and support a marketplace of iris-based applications based on standardized interoperable iris imagery. Our July 2013 IREX VI report[1] was motivated by a series of papers from the Bowyer group at University of Notre Dame (UND) that presented findings of "clear and consistent evidence of a template ageing effect that increases over time"[2]. This, and the UND press releases accompanying the publications, begat multiple instances of popular press coverage of the general form: "irises, rather than being stable over a lifetime, are susceptible to ageing effects that steadily change the appearance over time"[3]. The UND findings and the resulting press reports were contrary to the conventional view regarding the permanence of the iris as a biometric. Various government agencies asked us to resolve the contradiction.

IREX VI covered a) our new approach, the subject of the Bowyer-Ortiz criticisms, and b) our analysis of the UND images that underpinned the adverse headlines. The latter, which has *not* been disputed, showed that "template aging reported in the UND studies is largely due to systematic [pupil] dilation change over the collection period", a result with independent concurrence[4].

IREX VI set precedents by using logs from an operational system, NEXUS[5], and by applying linear mixed-effects regression to biometric comparison scores, an approach that has since been followed and extended by researchers at Michigan State for fingerprint[6] and face[7]. Our dataset's partitions include upto 521 thousand eyes, the largest at least two orders of magnitude larger than the sum of those in prior ageing studies. We qualified our ageing result as "provisional pending application of refined statistical techniques to larger and richer data sets, generalization to other recognition algorithms, better modeling of dilation and ... consideration of bilateral ageing in both eyes." We reported additional results at the International Biometrics Performance Conference[8].

As detailed later, UND interprets the International Standards Organization (ISO) standard[9] in a way at odds with the concept of ageing that most people observe in face ageing and that we modeled in IREX VI: a downward trend in similarity over time. UND, instead, regards even temporary changes to the iris as part of "template aging". The press and public officials who are not biometric experts, not surprisingly, interpreted the UND papers and press releases to be about permanence rather than the UND interpretation, thereby leading to the confusion we hoped to rectify.

A large part of the disagreement between the Notre Dame and NIST positions may be attributed to lack of application by Notre Dame of the G. E. P. Box maxim: "All models are wrong, some are useful". Notre Dame and NIST models for aging are both instances of "all models"; however, we believe that our models are more useful than the Notre Dame models for making decisions regarding the permanence of the iris as a biometric.

II. Comparison with prior work

Bowyer-Ortiz summarizes the prior ageing literature and claims that IREX VI conflicts with other research "all of which report observing a significant iris template aging effect." We reject this statement because Bowyer-Ortiz: a) omit some contrary work; b) levy an extraneous article; c) omit researchers' caveats; d) ignore poor image collection controls; and e) ignore inappropriate analyses.

Omitted references: Bowyer-Ortiz omit some contrarian results: Mehrota[4] reanalyzed the UND images noting false "rejections" in the UND papers "are caused by improper capture" concluding "While existing results are correct about increase in false rejection over time, we observe that it is primarily due to the presence of other covariates such as blur, noise, occlusion, and pupil dilation.", in agreement with IREX VI. Also they note that "only few samples ... are rejected and other samples of the same subject with similar time difference are accepted". Shchegrova[10] enrolled 236 images of 118 users of an operational iris access control system; with analysis of 8325 additional iris images collected over 900 days, she found that pupil size difference, iris size difference and overall image quality were strong, monotonic predictors of match score. She found that elapsed time is a predictor, but its partial dependence starts negative, turns positive, reaches a peak and turns back to zero over the course of the data collection. The paper concludes that ageing, using the permanence-based definition recommended by IREX-VI, is not seen. Gentric[11] showed no discernible adverse shift in HD histograms from the large-population logs of UK IRIS frequent traveler gates[12] covering up to seven years. Browning[13] used operational images and two algorithms to find accuracy variations that "did not follow a clear trend of greater dissimilarity as the time intervals increased that would be consistent with 'aging"'.

Extraneous reference: Fairhurst[14] report on whether age at the time of enrollment had an effect on match scores between enrollment and verification a relatively short and nearly fixed time later. The study is appropriately cross-sectional; it is not a longitudinal study of the effect of time elapsed between enrollment and recognition - the topic under discussion here.

Poor image collection controls: None of the papers cited by Bowyer-Ortiz adopted equipment and practices to provide strong control of presentation, dilation and other covariates. Hence, there are dilation and presentation variations that can vary systematically with time. Several papers acknowledge that dilation variation could be responsible for their observed temporal effects. We appreciate that adequate covariate measurement and control for a yet to be defined future analysis is difficult or impossible during the planning and execution of a data collection. Inadequacies may only become obvious in hindsight.

Inaccurate characterization of literature: Bowyer-Ortiz omit inconvenient qualifiers in several articles as described. Careful reading of these papers shows that a majority have opinions closer to IREX VI than to Bowyer-Ortiz. Further, four of the cited papers are from the UND group, using their datasets, collected using their protocols and analyzed using their processes, with one common author. There are variations from paper to paper, but they are hardly independent assessments. Tome-Gonzalez[15] demonstrated that genuine intra-session match scores have a significantly better distribution (more similar) than genuine inter-session (4, 8 and 12 weeks apart) match scores. This supports the common belief noted in ISO/IEC 19795-1(C.2.2)[9] that short term (minutes to hours) variability of a group of biometric samples is generally much smaller than for modestly longer intervals (days to weeks). Tome-Gonzalez note that inter-session distributions are not distinguishable: "... once a minimum time between samples has passed, error rates are not apparently increased." This statement is in full agreement with our position. Sazonova et. al.[16] showed a dependence of match score on elapsed time, and a 4.2% drop in true accept rate, but concluded with "it is possible ... that the quality metrics we have utilized are not adequately accounting for poor-quality images. ... different pupil dilations, the degradation from equipment, changes in data acquisition procedure, could contribute ... inter-actions between different factors may confound the true presence of iris template aging." Czajka[17] reports a significant difference in match scores between short term matches (same year) and those separated by 2 and 9 years. However, the effect from 2 to 9 years is smaller than that between 0 and 2 although the time span is 3.5 times longer. This is at odds with a continuing degradation associated with the familiar notion of ageing. In his conclusion, the author notes that dilation and other covariates may be contributing factors. Ellavarason and Rathgeb[18] used UND's data with a new set of algorithms (U. Salzburg Iris Toolkit) to find results similar to those of UND. They used UND's analysis, binning 2008-2009 and 2009-2010 pairs in a one year bin, thereby convolving an ageing effect with systematic dilation changes (Fig. 1) that occurred by year. In the masters thesis upon which the paper is based they state "It was found that there are various factors which account for high dis-similarity score in this data set such as pupil dilation, lens, illumination etc." and "...visual examination of iris images revealed that pupil dilation is not taken into account for comparison." This distinguishes Ellavarason from the Bowyer-Ortiz position and supports our analysis of the UND data.

In summary, we find only one group (UND) claiming compelling evidence of a template ageing effect that is *not* explained by covariates. Three more (Sazonova, Czajka, Ellavarason) find compelling evidence of a template ageing effect that may be explained by covariates or other effects. Six (NIST, Mehrota, Shchegrova, Gentric, Browning, Tome-Gonzalez) find a) scant evidence of a lack of permanence; and b) time variations in match scores that are largely accounted for by covariates. One group's work (Fairhurst) is on a separate topic. In addition, communications from two other entities (Iris ID and IrisGuard) support the results obtained by Gentric in separate analyses of other data sources.

Prior analysis: Bowyer-Ortiz fails to note that none of their cited papers address imbalance (disparity in the number of scores between subjects). Three papers of the papers [16], [19], [20] use simple linear regression that is contra-indicated in the opening pages of longitudinal analysis texts[21], [22]; for brief explanations, see Yoon[6] Fig. S2, its discussion, and the text for our Fig. 3 later.

III. Different goals and definitions

Bowyer-Ortiz' claims that a) "IREX VI did not attempt to study the same phenomenon studied in prior research" and b) "IREX VI conclusions contrast with previous research" cannot both be true. We hold that the cited prior papers had all been directed at checking Daugman's assertion that the iris texture is "immutable over a person's life"[23]. We are inclined to agree that the cited patent could not be definitive since everything - faces, fingers, irises, mountains, and monuments - ages. For IREX VI, the issue is how large an effect ageing is.

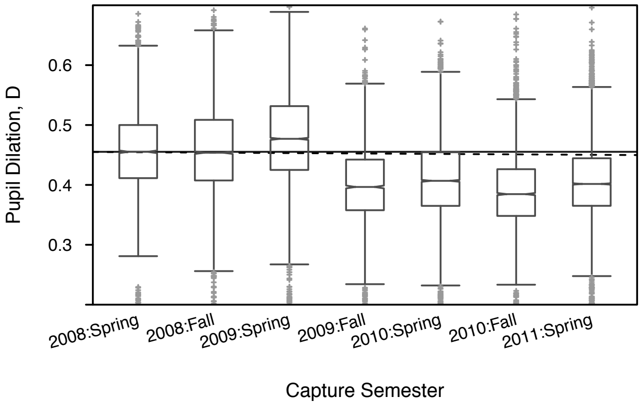

Fig. 1. Boxplots of pupil dilation over Notre Dame's ageing study[2]. The systematic shifts are contrasted with no change (grey line) and with that expected from lifelong pupil constriction (dashed line). The median dilation in Fall 2009 is 0.09 below that in Spring.

Bowyer-Ortiz Fig. 1 asserts that our definition of "iris aging" explicitly omits a variety of factors that would contribute to iris template aging "the increase in error rates caused by time-related changes in the biometric pattern, its presentation and the sensor". We agree with that statement: IREX VI is neither concerned with sensor ageing, nor particularly with non-monotonic temporal variations of human factors affecting presentation.

As construed by Bowyer-Ortiz, the ISO definition[9] allows arbitrary causes of time dependence to be classified as ageing: a blink would occlude the iris and impede recognition; diurnal variation in ambient lighting would affect face recognition[24]. Both of these are "time-related changes". Bowyer-Ortiz faults IREX VI for only focusing on the biometric pattern and not addressing "effect of sensor changes, environmental changes, behavioral changes, pupil dilation changes" as part of template ageing. We agree with Bowyer-Ortiz that their interpretation of the standard is different from our

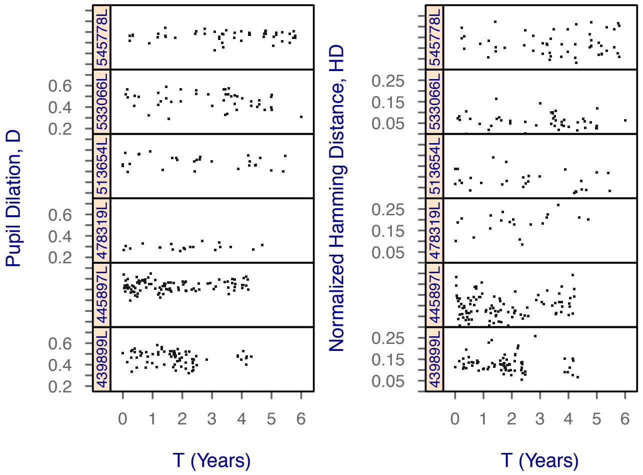

Fig. 2. Example trajectories of dilation (left) and HD (right) for six randomly selected left eyes presented over at least four years. Note imbalance across subjects and intra-subject variance. Note also inter-subject variance, some eyes' HDs are generally low, others high.

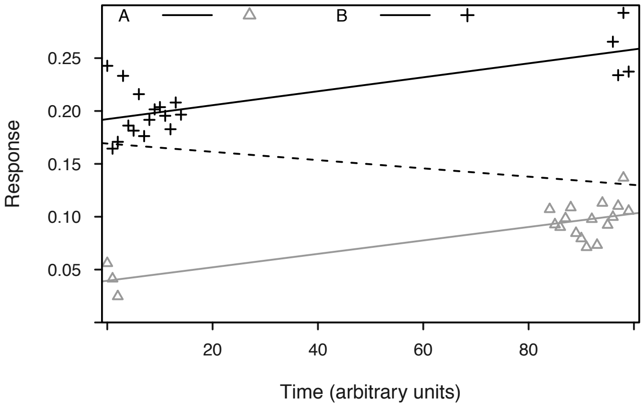

Fig. 3. Ordinary linear regression is inappropriate for longitudinal data, which is characterized by subject-specific variation, imbalance, and autocorrelation. The crosses, triangles and their fitted solid lines represent notional data from two subjects. Fits for both subjects have a positive slope. However, simple linear regression yields the dashed line with negative slope, because it does not heed identity information. The appropriate methodology, linear mixed-effects modeling, has subject-specific intercepts and trends. See Yoon[6] Fig. S2 for another ordinary least squares failure mode.

permanence-based definition of iris ageing in this context. We defined iris ageing around "irreversible changes to the healthy iris" in order to assess whether such hypothetical changes impugn iris as a long-term biometric. We did this by analogy to the familiar notion of face ageing as irreversible changes to the facial anatomy (jowls, creases etc.) that are known to erode both human and automated face recognition accuracy given sufficient time. IREX VI referred to "permanence" and perhaps should have adopted that word or "persistence" to be emphatic. We interpret all prior research, with the exception of UND, to be primarily directed at this irreversible component. For example, Ellavarason holds "Aging is a continuous process. Nothing or no one can advance or delay it. Usually it is slow and irreversible"[25].

In contrast, Bowyer-Ortiz state that ageing allows "Aunt Mary" to have "days when she is lucid and days when she is confused", advocating this as part of template ageing. Such oscillations cannot be aging, unless we allow for *negative* ageing. We're interested instead in monotonic trends.

In summary, we assert that most prior studies set out to assess permanence and to test Safir and Flom's[26] claim that iris will "permit ID for a substantial period", and Daugman's previously cited "immutability". There is an important place for analysis of sensor, environment and human factors variability in all biometrics. However, invoking the ISO standard to classify all of those sources of variability as ageing, leads to results which can mislead non-experts.

IV. Presence of false matches

Bowyer-Ortiz faults us for not noting the use of one-to-first search, a technique both ubiquitous and expeditious. Their point that it yields false matches is qualitatively correct, but both quantitatively irrelevant and incorrect. They estimate "hundreds of impostor scores" are present in the NEXUS data. However, their calculation is a factor of 14 too large. They allege a failure "to take into account ... the number of relative rotations" in iris matching so multiplied the Daugman false match rate calibration[27] by 14. However, as equations 8 to 11 and Fig. 6 in [27] show, this factor had already been included.

In arguing impostor scores are "important" they ignore rarity by declining to state them as a proportion of the 5.7 million in the NEXUS set. Correcting their "hundreds" to tens, the proportion is below 0.001%. That fraction should end the discussion; we note UND acknowledges the possibility of larger effects from ground truth errors in their collections[28] without noting the effect in their papers.

Finally false matches occur, by definition, due to similar iris templates. We expect them to be concentrated in certain individuals - Doddington's lambs[29]. In a mixed-effects regression, where every individual eye has its own HD vs. time trajectory (Fig. 2), the effect of impostor scores would affect only about 0.001% of eyes. Mixed-effects models therefore have additional immunity to impostor score contamination.

V. Truncated data

Section IV of Bowyer-Ortiz faults us for not applying truncated regression. Truncation arises in the NEXUS data for reasons of efficiency: HD computations are aborted as soon as the XOR count will give an HD above threshold. Bowyer-Ortiz incorrectly extract from [30] a truncation of 5% to 15%[30]. From the same table in the same report we find truncation to be from 3% to 8% for the cross-visit Panasonic-Panasonic pair used in the IREX VI ageing rate estimates. Whatever the number, match/non-match feedback to the subject during collection means that, in one sense, the score distribution is not truncated at all. Indeed, HD's above 0.27 are not seen, but the germane question is whether *individuals* are "truncated" from the data. If a specific individual did "age out" of the system, we would expect to see their trajectory revealing an ageing trend before their last successful attempt.

NEXUS data comes not from a collect-and-store process common in university data collections, but from a collect-recognize-decide feedback loop with matching essentially acting as a quality control enforcer. Most of the above-threshold events will be caused by aberrant presentations (such as eyes closed, blur, off-axis gaze) that can usually be remedied by immediate recapture of the iris. As discussed in the next section - second attempts usually produce a HD below 0.27. The only drops are people who give up and use an alternate mechanism. We are interested not in whether an iris is recognized at a particular time, but whether it *can* be. We're quantifying permanence, not human factors. A careless iris presentation by a weary traveler is immaterial if seconds later the iris is viable. Quality control of this sort was done (appropriately) by Baker[31], [32] in dropping about 90% of the images from the parent UND dataset[33].

Bowyer-Ortiz argue that our HD growth rate is underestimated with respect to theirs due to truncation. We don't dispute the qualitative effect, but we reject it quantitatively as follows: First, even though they cite their use of a mixed-effect software package, their methodology uses simple linear regression (teihr eq. (1)) that is inappropriate in that it handles neither imbalance (see Fig. 3) nor autocorrelation of the repeated measures. Second, even if their methodology was sound, their data shows that truncation at 0.27 leads to an underestimation of the ageing rate from $3.5\,10^{-5}$ day^{-1} to $3.0\,10^{-5}$ day^{-1}, i.e. just 15%. Applying that underestimation to our data, adds little to our stated uncertainty $(7.8 \pm 2.2)\,10^{-7}$ day^{-1}. For truncation to matter, there has to be an "aging" effect. In Notre Dame's data there is, due to systematic dilation changes seen in Fig. 1 and likely caused by uncontrolled ambient covariates[4], primarily illumination[1]. As a result, their "aging" estimate $3.5\,10^{-5}$ day^{-1} is 40 times higher than ours and should be discounted in any discussion of iris permanence.

Bowyer-Ortiz Section IV faults our "Failure to use a truncated regression technique" on the grounds that "OLS regression will not take into account the effect of truncation", failing to recognize that we never used ordinary least squares (OLS) - we did mixed-effects regression. Standard longitudinal analysis texts[21], [22] warn against OLS regression because it takes no account of a) imbalance, nor b) autocorrelation in the repeated measures. OLS assumes residuals are independent and homoscedastic, conditions that rarely hold in longitudinal data. In citing a particular truncated regression package, Bowyer-Ortiz recommends an approach that is patently inappropriate for longitudinal data. UND's use of regression[19], and the analyses in Bowyer-Ortiz (Sec. III-VII), all ignore individual-specific effects and do not correct for imbalance - the number of scores in [2] varies by an order of magnitude across subjects[1]. Their remarks here conceive of our dataset as being homogeneous but as shown in Fig. 2, there is substantial within- and between-individual variation. Some eyes give generally low HDs, others give higher values. Each has "noise" associated with the presentation of the eye and condition of the sensor. The mean responses differ, almost certainly due to variations in the qualities of the initial enrollment images. This heterogeneity is neatly handled by random-effects i.e. individual-specific offsets to population-wide (secular) intercepts and gradients.

Even though we and Bowyer-Ortiz demonstrated that the magnitude of the truncation effect in IREX-VI is small, we would take it into account if we could conveniently do so. However, to the best of our knowledge, the statistical community has not produced truncated regression techniques in mixed-effects settings. The choice here is between using a model (Bowyer-Ortiz) that can explicitly take into account a small truncation effect but can get the wrong sign for the overall effect (see Fig. 3) or using a model (IREX-VI) that takes proper account of the structure of the data, but which may have a small bias which, in IREX VI, is smaller than the estimated uncertainty in the parameter estimates. The reader can assess which model is more useful.

VI. Multiple attempts

Section V of Bowyer-Ortiz faults us for not including a covariate we didn't have. The attempt $(1, 2, \ldots)$ on which the subject's recognition succeeded only became available to NIST in 2015. Bowyer-Ortiz hypothesized that this is problematic after we informed them that a logged HD is the final result of a transaction involving one or more attempts (not just three as the they state), scores are logged if and only if the attempt matches below 0.27. UND views this operationally ubiquitous multiple-attempt reality as "replacing some (missing) above-threshold match scores with below-threshold scores" stating categorically that "this biases the estimated aging effect toward a lower-than-correct value."

We dismiss this issue on four grounds. First, as we stated previously, an iris permanence study should be interested in the state of the iris and whether it *can* be recognized. A poor quality first presentation necessitating a second attempt is immaterial if, seconds later, HD is low. Second, while we mostly agree with Bowyer-Ortiz that an ageing iris (or face) would need to be presented more often in eyes-fully-open second and third attempts, such aging would manifest itself as an increasing HD trend. By analogy, we're not interested in using car accident counts as a metric for speeding, we're interested in whether speeds are trending up. Ahead of the point when recognition cannot succeed at all, we'd expect an upward trend - exactly the effect quantified with our mixed-effects approach. Third, we note that regression is *never* conducted without missing covariates, and remains legitimate so long as the sampling does not introduce bias. Bowyer-Ortiz does not fault us for omission of constants such as eye color, race and height, nor for omission of time-varying covariates such as age, fatigue, hour, and airport. We didn't have any of this data, though all of it may well be material to recognition outcomes. Fourth, in any collection of biometric samples, there are poor quality images e.g. motion blurred, eyes closed, fingers improperly placed, faces occluded by hats, improper gaze or head pose etc. Whether these should be included, discarded, or discarded and recaptured is a matter of experimental design appropriate to the measurement goals. If we're counting poor quality images, we want to keep them. If the question is whether the subject biometric is changing over time, we posit that we can safely discard or discard/recapture. This is done implicitly in the NEXUS data, and explicitly by others[17] including UND[32].

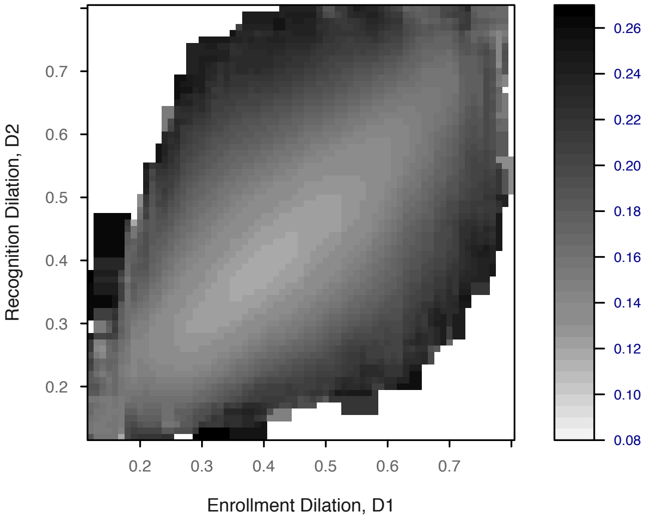

Fig. 4. Mean HD as a function of dilation in mated sample pairs. Two aspects of this figure demonstrate that dilation difference alone, $|D_1 - D_2|$ is an inadequate predictor of HD: a) elevated HDs at small and high dilations, and b) nonlinear gradients.

VII. Interpreting regression, collinearity, dilation

In Section VI Bowyer and Ortiz assert that we "underestimate the effect of aging" due to alleged "correlation among the independent variables" in our regression model. They restate our model but fail to note it as a mixed-effects model describing individual-specific trajectories (Fig. 2) about a secular (shared) population-aging rate. This omission betrays limited understanding of the approach and its elegant inclusion of identity information. Their remark that inclusion of a covariate and its square induces collinearity is simply wrong because they're represented orthonormally. We included quadratic dilation and dilation-difference terms because, as Fig. 4 shows, HD dependence on dilation resembles a quadratic bowl not a V-shaped valley: Dilation difference matters, but so do both small and large dilations. Higher degree polynomial terms are included ubiquitously in modeling texts[34].

Bowyer-Ortiz cites cross-sectional data[35] showing decades-long reduction in pupil size alleging collinearity with time. We reject it here as follows: First, pupil contraction is also accompanied by iris size[36] reductions which diminish the effect on dilation i.e. the ratio of pupil and iris size. Second, Bowyer-Ortiz declines to quantify the pupil size effect over the short durations of all published iris ageing studies (e.g. the dashed line in Fig. 1). Third, to make their point, they use a dataset known to have a collinear dependence built-in: Fig. 1 shows systematic increases in pupil dilation 2008-2009 and then larger decreases during 2009. In IREX VI we showed these systematic dilation variations explain the published UND template ageing effect. As noted above, Mehrota concurs[4]. In their critique, Bowyer-Ortiz run simple linear regression HD = B1 + B2 DilationDiff + B3 TimeLapse to show that collinearity reduces the B2 coefficient to "40% of" the value achieved by running just HD = B1 + B3 TimeLapse. In short, they introduce collinearity to show collinearity. Their result is not applicable to the IREX VI data.

Over the duration of our data dilation is essentialy a noise source: As is clear in Fig. 2 short term variance dominates long term trend. To formalize that, we quantify the variation of pupil dilation on short vs. long timescales using a two-level mixed-effects longitudinal model. Dilation of the i-th eye on the j-th occasion at time T_{ij} is modeled as the sum of four terms:

$$D_{ij} = \beta_{0i} \, + \, \beta_{1i} T_{ij} \, + \, \beta_{2i} \cos 2\pi(m_{ij} - 1)/12 \, + \, \epsilon_{ij} \qquad (1)$$

The first is an intercept, $\beta_{0i} = B_0 + b_{0i}$ consisting of a shared fixed effect, B_0, and an eye-specific random effect, b_{0i}. Even though we don't have age as a covariate, this allows for individuals to have age-specific intercepts per[35]. The second is a trend, $\beta_{1i} = B_1 + b_{1i}$ again consisting of population and individual components. The parameter B_1 is the rate we seek. The third is a purely fixed effect $\beta_{2i} = B_2$ modeling seasonal length-of-day effects at airports at 45^o latitude airports in month $m_{ij} = \{1 \ldots, 12\}$. As an aside, we posit cyclical dilation drives observed seasonal modulation of HD, shifted by about $\pi/2$, i.e. three months. The last term is the residual.

With the NEXUS dataset, we find an expected dilation trend of -0.0143 ± 0.0004 decade^{-1} buried in intra-eye noise ($\sigma = 0.047$) with substantial between-eye variation ($\sigma = 0.066$) due in part to age. These numbers allow us to dismiss Bowyer-Ortiz. Over three decades (0.047/0.0143) the life-long constriction process will result in dilation reducing by an amount equal to $\pm 1\sigma$ of the average eye's short term variation. Seasonal modulation ($B_2 = 0.00667 \pm 0.00005$) equates to almost five years of the long-term component. Stated another way, over the five year mean time lapse in the NEXUS data, the lifelong constriction component amounts to about 1/6 of one standard deviation of the day-to-day dilation variation. Thus dilation is mostly a "noise" source. It is barely collinear with time and its inclusion in our model supports better estimation of the ageing rate.

So does IREX VI have a collinearity problem, as Bowyer-Ortiz speculate? We don't think so given the above and the discussion here. We note first that collinearity exists in all longitudinal datasets (unless avoided by-design)[37] so the question becomes its degree. There, we are advantaged because the classic way to suppress collinearity is to add more data[37] (to enable the contributions of two covariates to be teased apart). We have very large amounts of data, small standard errors, and small independent variable correlations.

Second the standard indicator of collinearity, the variance inflation factor, VIF $= (1 - R^2)^{-1}$, is utterly benign. Obtained from a regression of two suspect covariates against each other, VIF becomes problematic if "unusually larger than 1.0"[38]. Williams suggests that VIFs much above 2.5 may be problematic in his Graduate Statistics II course at Notre Dame[39]. In our NEXUS partition, for dilation against elapsed time for each of the 43943 eyes partition, we see mean VIF of 1.017 with $< 0.01\%$ of eyes having VIF above 2.5. This is just a numerical statement of the flat dilation trajectories visible in Fig. 2. Similarly, the numbers for dilation difference and time are 1.014 with $< 0.01\%$ above 2.5.

Finally, even though dropping correlated covariates is not recommended, we indulge our critics by asking if the mixed-effects estimate of HD ageing rate changes if we exclude dilation? The answer is "It does!". On the NEXUS partition it goes from $(1 \pm 0.1)\, 10^{-6}$ day^{-1}, $p = 0$, with dilation, to statistically zero $(3 \pm 10)\, 10^{-8}$ day^{-1}, $p = 0.7$, without dilation. Including dilation in the model actually increases the magnitude of the reported ageing effect and usefully explains observed variance.

VIII. Operational Issues

Bowyer-Ortiz Section VII seems to complain that IREX VI did not provide a sufficiently comprehensive recommendation on how to deal with dilation in operational systems. We agree that IREX VI does not provide such recommendations, nor was it our intent to do so. We simply suggested that better dilation control is generally advantageous and point out some possibilities for consideration and discussion. Bowyer-Ortiz briefly discuss some of these options. We would be delighted to see and cite a paper from the Notre Dame group that provides comprehensive recommendations on how to appropriately control dilation in both laboratory and operational settings. Such a paper would be helpful to academics who are collecting data sets and to vendors who are fielding systems. We must correct Bowyer-Ortiz language in their Section VII - we did *not* endorse any particular method or vendor; our disclaimer[1](pg. 4) makes this clear.

IX. Nonlinear effects

Bowyer-Ortiz Section VIII, "Linear Regression Looks For Linear Effects", rather undersells linear regression by failing to note its "linear in the parameters" aspect[21](p.14) whereby it is possible to model nonlinear (e.g. seasonal or logarithmic) effects via any or all of: transformation of the dependent variable or the covariates; piecewise linear models; and polynomial terms.

Bowyer-Ortiz imply that we were remiss in not proceeding directly to nonlinear modeling because we cited arcus senilis and corneal shape change which they "suggest" are not linear "process[es] through adult life". We'd be the last to dispute that nonlinear approaches may ultimately be required given that some biological processes have variable and unknown rate parameters (e.g., the cause of presbyopia). However, if Bowyer-Ortiz are arguing that we were at fault for not doing nonlinear regression at the outset, then they seem not to appreciate the utility of mixed-effect models and these five points: First, we began with linear mixed-effects models because that's how longitudinal analysis conventionally proceeds[21] absent a mechanistic model[34](Chap. 6). Second, even if ageing is nonlinear, we hold that over a short enough duration everything is approximately linear - "graceful ageing". Our dataset extended to around nine years with a mean elapsed time below five years. As IREX VI stated, we ran a linear-time model because we have no basis for higher order terms. Third, we didn't have age data, so could not explore ageing as a function of age. However, random effects allow for exactly the kind of age-specific variation Bowyer-Ortiz imply. Importantly, individuals, young or old, can differ in their ageing rates and initial offsets. We draw our critics' attention to a) recent research[6] that includes age and elapsed time in mixed-effects approaches to biometric ageing, and b) standard texts' piecewise linear models[21]. Fourth, the Bowyer-Ortiz assertion that if an ageing rate coefficient "is near zero, one cannot conclude that there is no effect of TimeLapse on HD, only that there is no linear effect" is not relevant. It is simply a pedantic reminder that linear regression of, say, a sinusoidal variable over one period will produce a zero gradient estimate and "miss" the effect. However, the situation here is that if there is an ageing effect, by analogy with face ageing, it will result in a monotonic decrease in biometric similarity. Even if that were nonlinear, it would yield an aggregate increase in HD values and a non-zero regression gradient. The Bowyer-Ortiz statement is based on special cases that don't apply here and which require the existence of "negative aging" as noted earlier. Fifth, we note Bowyer and Ortiz are happy to cite linear-time regression models[16], [25], [35] and to run their own[19].

X. Summary

We have shown the Bowyer-Ortiz criticisms of IREX VI to be variously irrelevant, misinterpretations or qualitatively correct but quantitatively irrelevant. Their criticisms are rooted in mischaracterization of our approach, and misunderstanding of mixed-effects models. We differ with them in what ageing models and statistical tools are applicable to longitudinal biometric score data. We noted Box's maxim earlier - "All models are wrong, some are useful". The models in IREX VI are admittedly imperfect in the Box sense, but useful and they make use of standard practices for analysis of longitudinal data developed in other fields - even though these practices are relatively new to biometrics. UND's models and analyses of ageing have been, at best, misleading and in some cases simply wrong. The large "aging" effects claimed by UND in their series of papers, which led to headlines in the popular press, are the result of lack of control of ambient conditions during their data collections and do not represent changes in the underlying iris pattern. Their ageing effects ("153% increase in false non-match rate" over three years[2]) could have been realized in minutes via the same manipulation of the ambient illumination. As noted earlier, knowing what needs to be controlled is easy in hindsight; the publicly available UND data collections were and remain a significant service to the community. However, labeling the temporal effects reported in their papers based on that data as aging defies the commonly held conception of ageing. The UND "iris template aging" results to date should be deprecated in any discussion of iris permanence.

The Bowyer-Ortiz claim that IREX VI stands alone among papers on the topic of iris ageing is misleading at best and false at worst; a careful reading of the literature shows that, with the exception of the Notre Dame group, the prior literature is in basic agreement with IREX VI.

UND has obtained an updated version of the NEXUS logs from the Canadian Border Services Agency. We look forward to their investigations of the enhanced dataset. We encourage their progression from ordinary linear regression[19] toward individual-specific analyses and mixed-effects models.

DISCLOSURE

NIST has previously disbursed funds to UND and Michigan State University, although not for work on biometric ageing.

REFERENCES

[1] P. Grother, J. R. Matey, E. Tabassi, G. Quinn, and M. Chumakov, "IREX VI, temporal stability of iris recognition accuracy," NIST, Tech. Rep., July 2013, interagency Report 7948.

[2] S. P. Fenker and K. W. Bowyer, "Analysis of template aging in iris biometrics," in *Proc. CVPR Biometrics Workshop*, June 2012, pp. 45–51.

[3] D. Graham-Rowe, "Ageing eyes hinder biometric scans," *Nature/News*, May 25 2012. [Online]. Available: http://www.nature.com/news/ageing-eyes-hinder-biometric-scans-1.10722

[4] H. Mehrotra, M. Vatsa, R. Singh, and B. Majhi, "Does iris change over time?" *PLoS ONE*, vol. 8, no. 11, November 2013, e78333.

[5] NEXUS, "Canada border services agency," 2004-2013, http://www.cbsa-asfc.gc.ca/prog/nexus/menu-eng.html.

[6] S. Yoon and A. K. Jain, "Longitudinal study of fingerprint recognition," *Proc. National Academy of Sciences of the USA*, vol. 112, no. 28, pp. 8555–8560, 2015. [Online]. Available: http://www.pnas.org/content/112/28/8555.abstract

[7] L. Best-Rowden and A. K. Jain, "A longitudinal study of automatic face recognition," *Proc. IEEE ICB, to be published*, pp. 214–221, May 2015.

[8] P. Grother, J. R. Matey, G. W. Quinn, and E. Tabassi, "Quantifying biometric permanence using operational data: Longitudinal analysis of comparison scores," Proc. IBPC, April 2014, retrieved 06/06/2015. [Online]. Available: http://www.nist.gov/itl/iad/ig/ibpc2014.cfm

[9] T. Mansfield, *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*, JTC1/SC37/Working Group 5, August 2005, http://webstore.ansi.org.

[10] S. Shchegrova, "Analysis of iris stability over time using statistical regression modeling," in *Proc. Biometrics Consortium Conference*, September 2012.

[11] S. Gentric, "Iris aging and biometric performance," in *Proc. Biometrics Consortium Conference*, September 2013.

[12] UK IRIS, "Automated registered passenger entry to the uk," 2005-2012. [Online]. Available: www.ukba.homeoffice.gov.uk/customs-travel/EnteringtheUK/usingiris

[13] K. Browning and N. Orlans, "Biometric aging: Effects of aging on iris recognition," MITRE Corporation, Tech. Rep. 13-3472, September 2014.

[14] M. Fairhurst and M. Erbilek, "Analysis of physical ageing effects in iris biometrics," *IET Computer Vision*, vol. 5, no. 6, pp. 358–366, 2011, www.ietdl.org.

[15] P. Tomé-Gonzalez, F. Alonso-Fernandez, and J. Ortega-Garcia, "On the effects of time variability in iris recognition," in *Proceedings of the Biometrics: Theory, Applications, and Systems (BTAS)*, September 2008.

[16] N. Sazonova, F. Hua, X. Liu, J. Remus, A. Ross, L. Hornak, and S. Schuckers, "A study on quality-adjusted impact of time lapre on iris reocgnition," in *Proc. SPIE Biometric Technology for Human Identification IX*, vol. 8371B, April 2012.

[17] A. Czajka, "Template ageing in iris recognition," in *Biosignals - Conf. on 6th International Conference on Bio-Inspired Systems and Signal Processing*, no. 73, February 2013.

[18] E. Ellavarason and C. Rathgeb, "Template ageing in iris biometrics: An investigation of the nd-iris-template-ageing-2008-2010 database," Biometrics and Internet Security Research Group, CASED, Darmstadt, Germany, Tech. Rep. Nr. HDA-da/sec-2013-001, March 2013.

[19] E. Ortiz, K. W. Bowyer, and P. J. Flynn, "A linear regression analysis of the effects of age related pupil dilation change in iris biometrics," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS*, Arlington, VA, September 2013, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/BTAS.2013.6712687

[20] M. Trokielewicz, "Linear regression analysis of template aging in iris biometrics," in *Proc. Third IEEE International Workshop on Biometrics and Security*, Gjøvik, Norway, March 2015.

[21] G. Fitzmaurice, N. Laird, and J. Ware, *Applied Longitudinal Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2011. [Online]. Available: http://books.google.com/books?id=qOmxRtdNJpEC

[22] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press, March 2003.

[23] J. G. Daugman, "Biometric personal identification based on iris analysis," *U. S. Patent*, vol. 5,291,560 A, March 1994, filed 1991/07/15.

[24] Foto-Fahndung, "Face recognition as a search tool," Thaerstrasse 11, 65193, Wiesbaden, Germany, Tech. Rep., February 2007, bundeskriminamt (BKA). [Online]. Available: www.bka.de/kriminalwissenschaften/fotofahndung/pdf/fotofahndung_final_report.pdf

[25] E. Ellavarason, "Effects of ageing on iris biometric recognition," Masters thesis, Technical University of Denmark, Kongens Lyngby, June 2013.

[26] L. Flom and A. Safir, "Iris recognition system, u.s. patent 4,641,349," *USPTO*, 1987.

[27] J. Daugman, "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1927–1935, 2006.

[28] K. W. Bowyer and P. J. Flynn, "The nd-iris-0405 iris image dataset," University of Notre Dame, Tech. Rep., 06 2009.

[29] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *Proceedings of 5th International Conference of Spoken Language Processing*, ICSLP 98, Sydney, Australia, 1998, paper 608 on CD-ROM.

[30] M. Thieme, "Independent testing of iris recognition technology," International Biometric Group, Final Report, May 2005. [Online]. Available: https://www.hsdl.org/?view&did=464567

[31] S. Baker, K. W. Bowyer, and P. J. Flynn, "Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches," in *Proc. of International Conference on Biometrics*, 2009, pp. 1170–1179.

[32] S. Baker, K. W. Bowyer, P. J. Flynn, and P. J. Phillips, *Template Aging in Iris Biometrics: Evidence of Increased False Reject Rate in ICE 2006*. Springer, 2013, ch. 11, pp. 205–218.

[33] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Frvt 2006 and ice 2006 large-scale experimental results," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 5, pp. 831–846, 2010.

[34] J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[35] B. Winn, D. Whitaker, D. B. Elliott, and N. J. Phillips, "Factors affecting light-adpated pupil size in normal human subjects," *Investigative Ophthalmology & Visual Science*, vol. 35, no. 3, pp. 1132–1137, March 1994.

[36] L. A. Hall, C. Hunt, G. Young, and J. Wolffsohn, "Factors affecting corneoscleral topography," *Investigative Ophthalmology and Visual Science*, vol. 54, pp. 3691–3701, 2013.

[37] T. Baguley, *Serious Stats*. Palgrave Macmillan, 2012.

[38] E. R. Mansfield and B. P. Helms, "Detecting multicollinearity," *The American Statistician*, vol. 36, no. 3a, pp. 158–160, 1982.

[39] R. Williams, "Sociology 63993, Graduate Statistics II, Course Notes," Online, 2014, retrieved 6/12/2015. [Online]. Available: http://www3.nd.edu/ rwilliam/stats2/l11.pdf