NISTIR 8381 DRAFT SUPPLEMENT

Face Recognition Vendor Test (FRVT)

Part 7: Identification for Paperless Travel and Immigration

Patrick Grother
Austin Hom
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This document is a draft supplement of NIST Interagency Report 8381

2021/10/28

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8381 DRAFT SUPPLEMENT

Face Recognition Vendor Test (FRVT)

Part 7: Identification for Paperless Travel and Immigration

Patrick Grother
Austin Hom
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This document is a draft supplement of NIST Interagency Report 8381

October 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

INSTITUTIONAL REVIEW BOARD

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

ACKNOWLEDGMENTS

The authors are grateful for the support and collaboration of the U.S. Customs and Border Protection (CBP) component of the Department of Homeland Security.

Additionally we are indebted to staff at DHS' Science & Technology Directorate (S&T) and Office of Biometric Identity Management (OBIM) for discussions and image data that supports this work.

RELEASE NOTES

This report will be updated periodically with results for new algorithms, new analyses, and new datasets as they become available.

The report is open for comment - correspondence should be directed to frvt@nist.gov.

Executive Summary

We investigate the use of one-to-many facial recognition in airport transit settings in which travelers faces are matched against galleries of individuals expected to be present. We primarily consider the case where face recognition serves double-duty for access control (to an aircraft) and facilitation (of recording a visa-holders departure from a country). This is done in a paperless mode in which a boarding pass (something you have) is replaced with presentation of a biometric (something you are) to a camera, representing an implicit claim to be entitled to board. We describe how such systems can fail, discussing errors during gallery creation, photo capture at boarding, attack detection, and face matching. We discuss how errors might be estimated, citing relevant standards, and their consequences.

We quantify face matching errors by simulating departing flights, populating galleries with an airport ENTRY photo of 420 travelers, then measuring accuracy by running searches of EXIT photos. We repeat this with galleries populated with multiple photos per person, and with galleries as large as 42000, modelling the same concept of operations but at a centralized airport checkpoint. We report that accuracy varies greatly across algorithms, that use of multiple images per person reduces errors considerably, and that error rates when searching 42000-person galleries are often three times higher than in 420-person galleries, but still sometimes below 1%. We consider demographics, and note that for the more accurate algorithms, error rates are so low that accuracy variations across sex and race are insignificant. We include additionally a discussion of how our accuracy estimates might differ from those measured operationally due to by factors that we could not control, such as camera type and imaging environment.

Technical Summary

Background: One-to-many biometric search systems are discussed in their role of positive and negative identification - the former refers to the expectation that person in a probe sample is present in the database (as in access to an office) while the latter presumes the person is not (as in compulsive gamblers entering a casino). The distinction is useful because the applications differ in their tolerance for false negatives and false positives. This report addresses the positive use of one-to-many facial recognition in airport transit settings in which travelers' faces are matched against galleries of individuals expected to be present. We primarily consider the case where face recognition serves double-duty for access control (to an aircraft) and facilitation (of recording a visa-holder's exit).

In late 2018 the United States commenced a pilot of face-based confirmation of departure system in which passengers boarding an aircraft make cooperative presentations to a camera and the captured photos are immediately searched against a gallery comprised of photos of persons expected on the flight. This process is intended to biometrically bind the traveler to the departure. A positive biometric match in used two ways: First, by the airline, to grant access to the aircraft in lieu of a boarding-pass presentation; second, by passport control authorities to record the departure from the United States of in-scope passengers (e. g. visa holders), notionally replacing the long-standing airline manifest-based biographic process.

Overview: This report summarizes three NIST activities: First, to describe the biometric aspects of the traveler departure application and factors that are expected to affect its performance; second, to document results from running offline simulations in which recent accurate face recognition algorithms are applied to actual ENTRY and EXIT images with the goals of establishing a methodology, estimating accuracy, and exposing some factors that will affect those estimates; third to consider the use of face recognition at other airport touchpoints where higher populations are expected.

EXIT simulations: We simulate traveler EXIT by preparing 567 galleries each containing exactly 420 individuals representing the population expected on a flight. The individuals are not selected by age or sex. They are selected to have the same region of travel document (for example, South America or East Asia). We search each gallery with a fixed set of actual 132 931 EXIT photos using recent commercial one-to-many face recognition search engines. Notably we do not have camera, location and timestamp metadata so we cannot "replay" biometric boarding of actual flights. We

include this and other caveats in section 5.

Algorithms: Our EXIT simulations make use of one-to-many search algorithms submitted to NISTs ongoing Face Recognition Vendor Test between mid 2018 and April 2021. These algorithms are prototypes from the R&D laboratories of commercial developers of face recognition. These include two variants from the incumbent provider to the face matching facility used in the U.S., including the NEC-3 algorithm that was broadly the most accurate algorithm evaluated in 2018 as reported in NIST Interagency Report 8271 [1].

Images: This report makes use of images provided by DHS Office of Biometric Identity Management in May 2019. That collection is comprised of images and limited metadata indicating in which operation the data was collected e.g. airport-entry, pedestrian land entry, or exit. Some images were accompanied by metadata including date of capture, year of birth, sex, and country-of-birth[1]. From that database, this report uses 132 931 EXIT images of 128 384 individuals to search 567 air-ENTRY galleries each represented a departing flight[2]. Those galleries hold images drawn from the 825 976 airport ENTRY images of the 122 387 EXIT individuals who have a prior ENTRY image. The EXIT images were collected in 2018 and the first four months of 2019.

Other content: Section 1 discusses more general error sources and metrics relevant to EXIT and departure, putting matching results into the broader context of aircraft boarding. Section 2 guides readers toward different testing methodologies appropriate to answering a broader range of questions. Section 3 details our simulations and results. Section 4 considers use of one-to-many traveler verification systems (TVS) with a much larger population of N = 42 000 enrollees for use at other airport touchpoints. Importantly, section 5 discusses various reasons that would render the accuracy estimates in this report too high or too low.

Biometric results: We show that as many as 428 of 567 simulated flights each carrying 420 passengers can be boarded using one-to-many face recognition without any false negative errors - see Table 1 column 5. Stated in terms of error rates, this corresponds to at least 99.5% of travelers being able to board with a single presentation to a camera. This is attainable by enrolling a single prior ENTRY image in the galleries and using any of seven 2020-2021 face recognition algorithms - see Table 2 column 5.

For many travelers, multiple prior images can be enrolled in a gallery. Here, if we enroll an average of six prior air-ENTRY images, then the most accurate algorithm will now board 545 of 567 flights without any errors - see Table 1 column 4. Large gains are realized by all algorithms: Now at least 18 developers' algorithms are effective at boarding greater than 99.5% of travelers - see Table 2 column 3.

In 2007, U.S. legislation[3] specified that 97% of travelers exits should be verified. That requirement can be met with almost all of the algorithms tested here[4]. Note that there are various systematic reasons why such accuracy may not be achieved in practice - see section 5.

In test of late 2018 algorithms [1], the most accurate algorithms on large population mugshot searches were NEC-2 and NEC-3. They remain in the top five on that benchmark today. However, when matching lower quality EXIT to ENTRY images the algorithms are less accurate than a 2018 Microsoft algorithm and many other more recent algorithms. By taking 100 minus the miss percentages in Table 2, NEC-3 correctly identifies 98.7% of individuals enrolled with a single image and 99.0% of those enrolled with images from multiple prior encounters. For the most accurate algorithm, Visionlabs-10, these values are 99.9% and 100% respectively, corresponding to about a factor of 10 fewer errors than NEC-3. Note that NEC-3 is now more than two years old and we may assume NEC has since improved its capability.

[1] This metadata was vital to our 2019 quantification of demographic effects in NIST Interagency Report 8280. [2]
[2] The terms ENTRY and EXIT refer respectively to inbound and outbound border crossings to, in this case, the United States.
[3] See 8 U.S.C. 1187(c).
[4] We chose to run only recent and high-performing algorithms and also some widely used prior-generation algorithms. Many more algorithms have been entered into the 1:N search track of FRVT.

Our demonstration of considerably higher accuracy from newer algorithms is an existence proof that EXIT accuracy on operational images can be improved. Given the pace of developments associated with the industrial migration to various convolutional neural networks, it is incumbent on end-users to establish contractual provisions for technology refreshment, factoring in such quantities as speed, scalability, stability, and cost.

The accuracy values noted above correspond to correct identification of an individual here "correct" requires the algorithm to report the correct identity with a score above a set threshold. The threshold is set to limit false positives this is necessary to prevent illicit boarding of an aircraft in an access-control context, and to limit visa-holder's status indicator mistakes in an EXIT facilitation context. The false positive identification rate (FPIR) in this report is usually set to 1 in 3 333, i.e. the proportion of searches of people not entitled to board an aircraft who succeed in doing so. A false positive occurs when a photo from such a traveler matches any (random) gallery photo. The consequences of such events, and a more detailed discussion of security, appears in section 1.5. We also include figures showing the tradeoff of false negative and positive identification rates, noting that some algorithms can afford lower FPIR without greatly degrading accuracy. An FPIR of 1 in 3 333 would imply that a mismatch would occur once during the boarding of about eight flights (3 333/420) whether that is too frequent or to scarce is essentially policy issue informed by the error tradeoff characteristics of section 3.2.3 and the demographic dependencies given in sections 3.2.4 and 3.2.5.

Discussion: The report documents accuracy or small-gallery identification simulations showing a strong algorithm effect - accuracy is much improved with some algorithms versus others. This dominates two other main effects - first that more prior enrollment images for each enrollee improves accuracy and, second, that even a 100-fold population size increase degrades accuracy only modestly.

The report gives some information on demographic dependencies. Many algorithms give somewhat higher false negative rates on women compared to men. This is not true for, or has reduced magnitude, for the more accurate algorithms. With high accuracy, and with opportunities in real operations to make second identification attempts, these differentials are either small or can be remediated. The report also notes demographic dependence on false positive rates, particularly that women and people of certain nationalities, often East Asia, tend to give higher false positive identification rates. Again some algorithms are considerably superior to others in this respect. Note that security context matters: In particular that passive non-mate, and active attack, presentations will be very small percentages of all attempts.

The accuracy estimates in this report are just that, estimates. Section 5 notes several factors that would drive accuracy higher or lower. Primary among those is that we cant be sure how well the images we possess represent the actual paired ENTRY galleries and their EXIT photos. A passport control authority has two complementary options for improving on our estimates: First is to run exhaustive clipboard style operational tests; second is to provide NIST or some other laboratory with a) actual images, and b) the operational algorithm. This latter option had been planned in 2019 but was derailed for several reasons, including the COVID pandemic.

ALGORITHM			NUM ZERO FALSE NEGATIVE SIMULATIONS		
			$N = 420$	$N = 420$	$N = 42000$
#	NAME	DATE	$k \geq 1$	$k = 1$	$k = 1$
1	VISIONLABS-010	2021-02-05	[1]545	[3]428	[6]177
2	SENSETIME-006	2021-07-26	[2]543	[1]475	[2]282
3	IDEMIA-008	2021-03-15	[3]536	[4]422	[4]215
4	VISIONLABS-009	2020-08-04	[4]533	[5]406	[9]125
5	NTECHLAB-010	2021-06-24	[5]533	[7]389	[8]126
6	CUBOX-000	2021-08-24	[6]533	[2]434	[1]304
7	CLOUDWALK-HR-000	2021-02-10	[7]528	[6]393	[3]265
8	DEEPGLINT-001	2020-07-23	[8]519	[9]336	[7]153
9	CANON-CIB-000	2020-10-19	[9]518	[11]307	[19]19
10	XFORWARDAI-002	2021-05-24	[10]513	[8]348	[5]198
11	XFORWARDAI-001	2021-01-21	[11]513	[10]309	[11]113
12	INCODE-005	2021-07-29	[12]512	[12]299	[20]18
13	IREX-000	2021-02-09	[13]511	[14]261	[27]2
14	CYBERLINK-003	2021-01-08	[14]504	[13]287	[16]36
15	PARAVISION-007	2021-02-01	[15]490	[15]237	[10]124
16	TEVIAN-006	2021-04-16	[16]480	[16]235	[21]17
17	TRUEFACE-000	2021-01-27	[17]476	[24]154	[23]4
18	NEUROTECHNOLOGY-008	2021-03-26	[18]470	[21]169	[25]2
19	RANKONE-011	2021-08-27	[19]464	[20]173	[29]1
20	COGENT-004	2021-02-10	[20]454	[19]182	[22]10
21	PARAVISION-005	2019-12-11	[21]453	[22]156	[14]72
22	NTECHLAB-008	2020-01-06	[22]451	[26]125	[31]1
23	PIXELALL-004	2020-07-02	[23]435	[25]146	[32]0
24	TECH5-002	2021-04-07	[24]416	[27]110	[26]2
25	DERMALOG-008	2021-01-25	[25]382	[30]71	[39]0
26	IDEMIA-007	2020-01-17	[26]374	[31]66	[33]0
27	MICROSOFT-006	2018-10-29	[27]361	[23]155	[24]3
28	SENSETIME-005	2020-12-17	[28]319	[17]233	[12]99
29	SENSETIME-004	2020-08-10	[29]316	[18]208	[13]96
30	RANKONE-010	2020-11-05	[30]300	[29]76	[34]0
31	COGNITEC-005	2021-07-30	[31]239	[35]24	[28]1
32	RANKONE-009	2020-06-26	[32]203	[34]38	[30]1
33	COGNITEC-004	2021-03-08	[33]201	[37]11	[37]0
34	NEC-002	2018-10-30	[34]111	[33]65	[17]32
35	NEC-003	2018-10-30	[35]111	[32]66	[18]30
36	NEUROTECHNOLOGY-007	2019-10-03	[36]90	[36]21	[35]0
37	IDEMIA-004	2018-06-30	[37]3	[38]0	[38]0
38	NEC-000	2018-06-21	[38]0	[39]0	[36]0
39	NEC-004	2021-07-16	[39]0	[28]78	[15]56

Table 1: **Number of simulations (out of 567) completed without errors.** *The second row N values give the number of individuals enrolled in each gallery. The 420 person galleries represent aircraft boarding; the 42000 case represents a airport security line where many more people are expected. The third row k values give the number of images of each enrollee in each gallery.*
The second and third columns identify the algorithm and the date it was submitted to NIST. The remaining columns give the number of simulations, out of 567, for which all 420 travelers boarded the flight (cols. 4, 5), or passed the checkpoint (column 6), without experiencing a false negative. Higher values are better, and the table is sorted on the first results column. The threshold is set so that only a fraction, 0.0003, of non-mated searches would return any match. The shaded cells indicate the three most accurate algorithms for that trial.

ALGORITHM			PERCENT TRAVELERS NOT MATCHED		
			$N = 420$	$N = 420$	$N = 42000$
#	NAME	DATE	$k \geq 1$	$k = 1$	$k = 1$
1	VISIONLABS-010	2021-02-05	[1]0.02	[3]0.13	[6]0.61
2	SENSETIME-006	2021-07-26	[2]0.02	[1]0.08	[2]0.35
3	IDEMIA-008	2021-03-15	[3]0.02	[4]0.15	[4]0.49
4	VISIONLABS-009	2020-08-04	[4]0.03	[5]0.16	[8]0.74
5	CUBOX-000	2021-08-24	[5]0.03	[2]0.13	[1]0.31
6	NTECHLAB-010	2021-06-24	[6]0.03	[7]0.18	[9]0.77
7	CLOUDWALK-HR-000	2021-02-10	[7]0.03	[6]0.18	[3]0.43
8	DEEPGLINT-001	2020-07-23	[8]0.04	[9]0.24	[10]0.80
9	CANON-CIB-000	2020-10-19	[9]0.04	[11]0.30	[19]1.79
10	INCODE-005	2021-07-29	[10]0.05	[12]0.31	[21]1.92
11	XFORWARDAI-001	2021-01-21	[11]0.05	[10]0.28	[11]0.81
12	XFORWARDAI-002	2021-05-24	[12]0.05	[8]0.23	[5]0.51
13	IREX-000	2021-02-09	[13]0.05	[14]0.39	[28]3.91
14	CYBERLINK-003	2021-01-08	[14]0.05	[13]0.32	[16]1.44
15	PARAVISION-007	2021-02-01	[15]0.07	[15]0.41	[7]0.72
16	TEVIAN-006	2021-04-16	[16]0.08	[16]0.41	[20]1.85
17	TRUEFACE-000	2021-01-27	[17]0.08	[23]0.66	[25]3.72
18	NEUROTECHNOLOGY-008	2021-03-26	[18]0.08	[20]0.59	[29]4.12
19	RANKONE-011	2021-08-27	[19]0.09	[19]0.59	[26]3.83
20	PARAVISION-005	2019-12-11	[20]0.10	[22]0.62	[14]1.04
21	NTECHLAB-008	2020-01-06	[21]0.11	[26]0.81	[30]4.52
22	COGENT-004	2021-02-10	[22]0.11	[21]0.59	[22]2.34
23	PIXELALL-004	2020-07-02	[23]0.12	[24]0.69	[27]3.88
24	TECH5-002	2021-04-07	[24]0.14	[27]0.86	[32]5.21
25	DERMALOG-008	2021-01-25	[25]0.19	[28]1.04	[34]6.39
26	IDEMIA-007	2020-01-17	[26]0.19	[30]1.12	[31]5.19
27	MICROSOFT-006	2018-10-29	[27]0.23	[25]0.71	[23]3.21
28	SENSETIME-005	2020-12-17	[28]0.28	[17]0.45	[12]0.85
29	SENSETIME-004	2020-08-10	[29]0.29	[18]0.50	[13]0.89
30	RANKONE-010	2020-11-05	[30]0.31	[29]1.06	[33]5.71
31	COGNITEC-005	2021-07-30	[31]0.43	[37]2.18	[24]3.41
32	COGNITEC-004	2021-03-08	[32]0.49	[36]2.18	[36]9.20
33	RANKONE-009	2020-06-26	[33]0.52	[34]1.52	[35]7.85
34	NEC-002	2018-10-30	[34]0.99	[32]1.29	[17]1.61
35	NEC-003	2018-10-30	[35]0.99	[33]1.29	[18]1.78
36	NEUROTECHNOLOGY-007	2019-10-03	[36]1.02	[35]2.02	[38]31.93
37	IDEMIA-004	2018-06-30	[37]4.96	[38]8.13	[37]17.81
38	NEC-000	2018-06-21	[38]15.41	[39]18.85	[39]91.97
39	NEC-004	2021-07-16	[39]27.73	[31]1.19	[15]1.34

Table 2: **False negative rates by gallery size and number of enrolled images per person.** *The second row N values give the number of individuals enrolled in each gallery. The 420 person galleries represent aircraft boarding; the 42000 case represents a airport security line where many more people are expected. The third row k values give the number of images of each enrollee in each gallery.*
The second and third columns identify the algorithm and the date it was submitted to NIST. The remaining columns give false negative identification "miss" rates i.e. the proportion of travelers not matched to their gallery photo(s), expressed as a percentage. Lower values are better, and the table is sorted on the first results column. The superscripts give the rank of the algorithm for that column. The threshold is set so that only a fraction, 0.0003, of non-mated searches would return any match. The shaded cells indicate the three most accurate algorithms for that trial.

1 Errors and Their Consequences in Biometric Exit

The following subsection describe mechanisms by which an EXIT system, as comprised, makes errors. We distinguish biometric errors (from cameras and algorithms) from operational issues deriving from business processes.

1.1 Failure to Enroll

Nature: In the context of TVS manifest-driven gallery construction, some individuals who are legitimately booked on an aircraft will not be enrolled in the face recognition gallery. This number will usually be zero but could be non-zero for several reasons, among them:

1. Absence of historical photo. For various policy-related issues a PCA may not have a prior photo - these could include first-time visitors, foreign passport holders born in the country, and bilateral trade-related visa exemptions. In such cases a PCA might legitimately have no ENTRY record. This circumstance might be termed an *operational failure to enroll*.

   Measurement: A PCA can estimate the prevalence of missing enrollments by cross-referencing airline manifests and the lack of prior reference photos. This estimate will include instances of 2 below.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

2. Biographic errors. It is possible that the manifest provided to the PCA by the air carriers includes biographic errors from well understood sources such as recent marriage and change of name, and typographical errors.

   Measurement: A PCA can estimate the prevalence of missing enrollments by cross-referencing airline manifests and the lack of prior reference photos. This estimate will include instances of 1 above.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

3. Poor image quality. It is possible the photographs that a PCA has on an individual are of poor enough quality that the TVS feature extraction software fails to produce a template from the photograph. This could occur because the face detector fails to find the face, or because the software deems the photo to be of low utility to their downstream recognition engine so, electively, does not produce a template. Such outcomes would constitute *biometric failures to enroll*.

   Measurement: A PCA can estimate algorithm enrollment failures by direct analysis of TVS logs.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

1.2 Failure to Capture

Nature: During aircraft boarding TVS never receives photos of some travelers for at least two reasons:

1. Camera failure: Some cameras might fail to trigger and take a photograph. This can occur due to failed face detection (e.g. due to sunglasses, or subject not being in the field-of-view), or because an on-board quality algorithm deemed the captured photograph of insufficient utility, or due to some system fault of the kind remedied

by rebooting the system. During observations at various airports in June 2019, some cameras would not trigger; others would trigger only after the subject disengaged by moving away, and then re-engaged.

Measurement: We can put an upper bound on the frequency of such events by subtracting the number of people verified from the number of people on the manifest. This quantity will include outright recognition failures too. This estimate will include people who never appeared before the camera (e.g. because the airline allowed traditional paper-based boarding).

2. Airline operations: An operational source of "failure to capture" can be that airline staff might redirect the traveler to some human-adjudicated boarding process such as the traditional passport or boarding-pass based biographic confirmation. This could occur a) because the staff perceive the traveler has had difficulty, or b) that they will have difficulty (e.g. because theyre too tall or short), or c) simply because the airline staff are trying to expedite boarding by using the biometric process and the biographic process.

   Measurement: Such events can only be documented by observation, most readily human observation, but also via some automated supervisor or logging system.

Consequences: For an in-scope traveler the consequence will be that EXIT will only be recorded biographically according to the information used in forming the passenger manifest this is essentially the legacy biographic process. An immediate operational consequence is that the passenger will have to be processed manually (by airline) staff. Downstream, this may cause the PCA to perform overstay inquiries.

1.3 Failure to Extract Features

Nature: It is possible the photographs that the PCA has on an individual are of poor enough quality that the TVS feature extraction software fails to produce a template from the photograph. This can occur during gallery construction or during EXIT operations.

Measurement: Such events can be measured from algorithm logs such as those produced in FRVT, and likely by operational systems.

Consequences: If TVS fails to extract features during EXIT, the travelers boarding attempt will be rejected, possibly silently. He or she may make a second attempt, perhaps after being prompted. In June 2019 observation of boarding, the author noticed airline staff directing passengers to the gate-agent biographic process. This would likely lead to the PCA having to revert to its reliance on biographic recording of EXIT.

1.4 False Negative During Identification

Nature: In a positive identification application like EXIT, the one-to-many search algorithm generally grants access if the rank-one (i.e. highest-scoring) candidate has a score above threshold. The identity of the person in the live photo is taken to be that returned by the system even if it is incorrect. From a testing perspective, an error occurs if the rank-1 candidate is of the wrong identity or has score below threshold. This gives us the following performance metric, the false negative identification rate (FNIR):

$$\text{FNIR}(N,T) = \frac{\text{Num. searches where top-scoring candidate has wrong ID or score below threshold}}{\text{Number of searches conducted}} \quad (1)$$

This definition automatically incorporates failure to extract feature events as they wont return high-scoring candidates. The dependence on gallery size, N, and threshold, T, are present as they are design choices affecting FNIR. For an

audience who likes to think in terms of accuracy or hit rates, we can convert the "miss rate" or Eq. 1 to True Positive Identification Rate using TPIR = 1 FNIR, so a 3% FNIR becomes 97% TPIR. However, that definition is nave in that it assumes every traveler was photographed. It ignores instances of failure-to-capture, and also cases where travelers are photographed, not matched, and then make further attempts. Then an operational definition of false negative identification rate is

$$\text{FNIR}(N,T) = \frac{\text{Num. travelers who are not matched to the correct ID in one or more presentations to the camera}}{\text{Number of travelers}} \quad (2)$$

The two measures would be equivalent if each traveler executes just once search. To the extent that is true, our Equation 1 estimates in this report will approximate Equation 2. We use the 1 thoughout this report. We discuss in section 5, factors that can make our estimates too high or too low.

Measurement: In this report, we dont have insight into the transactional nature of aircraft boarding, with failed captures or failed searches. Instead all we see are images that can be used in simulations of boarding. For measuring duration of boarding, and quantities such as the number of travelers who need to make further presentations, an operational observational test is most appropriate. Many aspects may be measurable in scenario tests in which passengers and airline staff model the actual target boarding process.

Consequences: False negatives will usually be resolved by biographic and human visual biometric inspection by airline staff. For an in-scope traveler the consequence will be that EXIT will only be recorded biographically according to the information used in forming the passenger manifest this is essentially the legacy biographic process. Downstream, this may cause the PCA to perform overstay inquiries. The PCA would possess an aircraft boarding photo, but one that is not bound to an identity such images are provided to PCA staff monitoring a flight departure.

1.5 False Positive During Identification

Nature: False positives occur when images of two people are erroneously associated. In biometric EXIT there are three kinds of false positive:

- First is the **in-gallery false positive** in which a legitimately enrolled traveler matches the wrong identity. Such a possibility necessarily implies that the correct identity would be displaced from the rank-1 position on the candidate list, usually to rank 2. That list is a data structure internal to the particular TVS and is not typically presented to airline staff or anyone else. Depending on how the system is built, an in-gallery false positive may result in a false negative for the correct passenger if he or she boards later in the process. Such errors were observed by the author in June 2019 during visits to observe the boarding process in five different airports.

  Measurement: The in-gallery false positive rate is not currently defined in performance testing standards as it is approximately the proportion of mated searches yielding the mate at rank 2 or higher. Such an outcome would most often occur because the search imagery is of poor quality, but could occur if the enrolled imagery was poor. Formal measurement can be achieved by careful online observation of the boarding process. Error rates can be estimated approximately from recognition logs by counting instances of a passenger apparently boarding the plane twice once legitimately as themselves and secondly when another traveler incorrectly matched their identity.

  Consequence: Such errors will likely be resolved by airline staff, who may become familiar with such an event.

- Second is the **incorrect acceptance** of people who are not in the gallery and not expected on the departing flight.

This population includes travelers who mistakenly arrive at the wrong gate[5] without subversive intent. The frequency of occurrence is usually stated by the False Positive Identification Rate (FPIR). FPIR is the primary security-related parameter in a one-to-many access control system. Its value is chosen by a system owner to target security objectives and is implemented by setting the system threshold according to some calibration[6].

Measurement: Such errors were observed by the author in June 2019 when airline staff in the gate area were accidentally captured by the camera and incorrectly matched to an actual passenger. While this kind of error could be measured by making in-person attempts, this approach does not scale. An offline approach in which images are matched after-the-fact affords more precise FPIR estimates this report takes just this approach.

Consequence: The consequence for the airline is potentially a stowaway. However, airlines usually count passenger totals and may thereby be able to detect such events. While there is little consequence for the PCA's EXIT processing, these events, if undetected, could cause erroneous updates to the PCA's systems, undermining integrity.

- Third is a false positive from someone who is illicitly trying to gain access. This category would include stowaways and potentially visa overstayers.

> **Passive vs. active attack**
>
> False match rates usually express the likelihood that a face recognition algorithm will compare two photographs and return a high score from two individuals who are selected entirely randomly, or perhaps with the restriction that they have the same demographics such as age, sex, and race.
>
> However, if someone makes more deliberate efforts to impersonate an identity e.g. via cosmetics or wearing a face mask, then additional algorithms must be employed to detect the presentation attack (PA). To succeed an attacker must defeat the PAD subsystem, if installed and enabled, AND match the intended identity see section 1.6

  - **Casual attack:** If someone is making a low-effort attack for example as a stowaway they might rely on matching any identity essentially fortuitously, and then hoping the airline staff does not notice nor take steps to resolve the match. A second intent here would be to fake someones departure from a country. This possibility - to overstay a visa by sending a confederate to verify a particular identity - is notable in that it would be difficult for an overstayer to select a confederate who would match the *particular* identity in a biometric search In this respect a one-to-many system where there is no claim to an identity is more secure to passive attack. However, the security context is that such a system is prone to circumvention attack: a confederate failing to match an enrolled identity might appeal to airline staff who would make biographic or visual biometric efforts to verify the person, with the likely outcome that passenger would be allowed to board.
  - **Active attack:** An overstay attempt would be much more successful if the confederate actively impersonates the visa-holder. This could be achieved using a presentation attack instrument such as a face mask.

    Measurement: Vulnerability to active attack could be demonstrated via "red-team" presentations to the operational system. More formal quantification of the vulnerabilities is best conducted in laboratory trials using identical equipment to that used in the operation. Each approach will require controlled, defined and

[5] This can occur because of a gate change, or because someone goes to the wrong gate. The author, for example, has accidentally tried paper-based boarding at the adjacent gate on several occasions.

[6] Threshold calibration is an imprecise process because FPIR often depends on demographics and image quality related properties. A threshold is set starting with vendor recommendation and refined using offline tests (such as FRVT) or empirical instrumentation and tests or logging of the operational system.

repeatable production of presentation artefacts (masks, cosmetics etc.). The metrics relevant to this kind of attack are standardized see section 2.2.

- **Comparison with existing paper-based boarding:** Attacks on non-biometric paper-based departure systems are possible also: A stowaway could find, or steal, a boarding pass. A confederate seeking to depart for a visa-overstayer would only have to present a boarding pass and possibly a cursory inspection by the airline staff of the passport. In these cases, a biometric system, if used and not circumvented, will improve security compared over the legacy process.
- **Consequences:** For an IA, a successful impersonation attack would likely produce an undetected overstay. The attack assumes the confederate either does not need or want to return to the United States or could do so using other documents. There are no consequences for the airline.

1.6 Presentation Attack Detection Metrics

The ISO/IEC 30107-3 standard establishes the metric Impostor Attack Presentation Match rate (IAPMR) which expresses the proportion of attackers who both defeat the PA detection software AND match the correct identity. That metric is appropriate to access, say, to a mobile phone. In one-to-many processing such as paper-less EXIT, a traveler would have to defeat the PAD and match the specific intended enrollment.

1.7 Demographic Differentials

Biometrics generally give different error rates for different populations. For example, fingerprints are known to give higher false negative rates in the very young and the elderly[7]. NIST Interagency Report 8280 [2] documented error rate differentials for face recognition examining the effect of sex, age and race on accuracy of many commercial algorithms. That report made an important distinction between differentials in false negative and false positive error rates, the former affecting how well a single individual is not matched as him or herself, the latter affecting how often two individuals are erroneously associated. The consequences of such errors, and differentials in their rate of occurrence, are very different. We include visualizations of false negative differentials in section 3.2.1 and false positives demographic differentials in sections 3.2.4 and 3.2.5.

[7] In the young, typical contact sensors have inadequate resolution to resolve the fine friction ridge structure. In the elderly the factors include inelasticity of the skin and inability to present flat impressions e.g. due to arthritis.

2 Operational Questions

2.1 Context

This report gives extensive documentation of biometric identification performance. However larger questions exist, and core biometric performance statements only inform answers to those questions. For example,

- An airline might ask "which camera and boarding solution should we procure?" this report is silent on that because we would at least need to know what cameras were used for collecting the data, and this is not information we have. Dedicated laboratory tests of camera equipment[8] are appropriate to such tests.
- An airline might ask "what is the proportion of passengers being referred to gate agents" such a quantity could be approximately estimated from TVS logs, but is more precisely answered only by observation of the operational system.
- A security analyst might ask "what is the chance on an active impersonation attack succeeding" this question can be addressed potentially by laboratory trials if the fielded system can be copied and if access access to the TVS recognition engine is granted. It may be easier to conduct operational "red team" trials with an appropriately motivated staff. Active attacks (e.g. using face masks) are not the fault of the recognition algorithm per-se, but are enabled by lack of (or use of poor) presentation attack detection algorithms[9].
- A policy maker might ask "is biometrics better than biographic matching for overstay detection?" we cant address that without biographic data and extant biographic matching algorithms.

2.2 Standardized Tests

Since 2003, there have been significant worldwide investments in supporting development of biometrics performance testing and reporting standards in the ISO/IEC JTC 1 Subcommittee 37. That body develops very well vetted consensus standards in working groups (WGs) dedicated to vocabulary (WG1), interfaces (WG2), data interchange and image quality (WG3), application aspects including face-aware capture devices (WG4), performance testing and reporting (WG5) and societal issues (WG6). Table 3 lists standards that may be valuable in the measurement of performance in a PCA's ENTRY-EXIT processes.

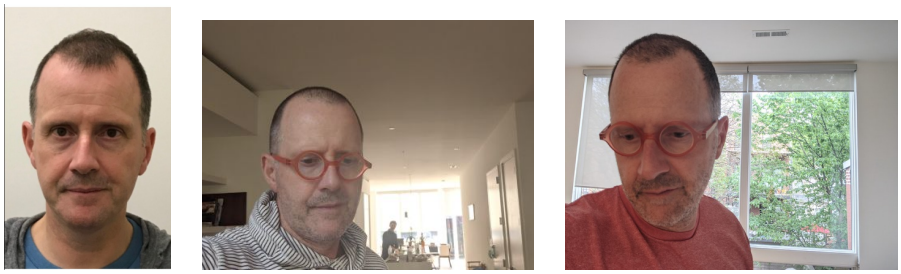
There are a number of other testing standards supporting other domains of use.

[8] See the scenario tests conducted at the Maryland Test Facility, for example.

[9] PAD approaches have advanced in recent years, both in software and hardware. However, their use will often increases false negatives because they sometimes erroneously flag a bona-fide presentation. Their use may be more appropriate on inbound arrival processing (ENTRY).

Table 3: Testing standards supporting performance measurement in ENTRY-EXIT

Number	Title	Relevance
ISO/IEC 19795-1	Principles and Framework	This foundational document establishes requirements on all biometric tests regarding design of tests of enrollment, verification and identification, and how to put uncertainty estimates on measured error rates.
ISO/IEC 19795-2	Technology and Scenario Testing	Regulates two kinds of in-vitro test: Technology tests which are most often offline sample comparison and search tests such as those documented herein, and scenario tests that are usually human-in-the-loop laboratory tests intended to mimic operational systems.
ISO/IEC 19795-3	Environmental Aspects	A technical report guiding testing and reporting in the presence of environmental variations such as humidity and illumination
ISO/IEC 19794-4	Interoperability Testing	Relevant to tests where components of a system, possibly from different manufacturers must produce and consume standardized data, for example cameras must produce images that will be consumed by remote recognition algorithms.
ISO/IEC 19795-6	Operational Testing	Establishes requirements on in-situ tests, where identity ground truth is not necessarily known, and where the act of measuring accuracy or duration can potentially disturb the estimates. This kind of test is advantaged by considering the actual system on its native population in its native environment. These aspects are often material and difficult to approximate in lab tests.
ISO/IEC 30107-3	Presentation Attack Detection	This standard regulates tests of PAD components and PAD-enabled systems and gives detailed guidance on measuring and naming of error rates that are available for various levels of logging and instrumentation.
ISO/IEC 19795-10	Demographic dependence	This standard (2020-11) is in the early stages of development. It will establish requirements on various kinds of tests intended to measure demographic differentials in biometric devices, algorithm and systems.



a) Visa or Passport | b) Immigration Lane ENTRY | Departure Point EXIT

Figure 1: Image (a) is representative of passport-like data that would ordinarily be available to a PCA's TVS from all in-scope travelers and citizens. However, such images were not available for the trials conducted here. The remaining images have size 240x240 pixels and are representative of some poorer quality ENTRY images: Image (b) is typical of ENTRY photos in that it has has non-frontal pose, and strong background illumination reducing contrast on the subjects face. Image (c) is typical of EXIT photos in that it exhibits some close-range distortion, mild non-frontal pose arising from "don't wait for frontal presentation" fast-capture and adverse background lighting. contrast.

3 Simulation and Results

3.1 Air-Exit Simulations

We simulate biometric EXIT by running simulations using archived images as follows.

1. We form a departing flight by placing ENTRY images from N = 420 individuals into a gallery. We use 420 because that number is reasonable for a large commercial twin-aisle jet such as the Boeing 777 or the Airbus A380[10]. The exact gallery size is not that important because accuracy is an insensitive function of N. We later increase the population size to 42 000 to simulate an airport security checkpoint, for example.

2. We populate an EXIT gallery in two ways.

   (a) First, with one ENTRY image per person.

   (b) Second with a multiple such images, the average is about 6, with some variance per individual.

   While it is common practice to populate the gallery with images from all prior encounters of a person[11], we include the one-image case to show "worst-case" accuracy i.e. that expected when only one prior encounter is available. We include results for single- and multiple-image enrollment in sections 3.2.1 and 3.2.2 respectively[12].

3. We populate a gallery with individuals from the same region of the world. We do this for two reasons: As discussed in section 5, we list various factors that will push our error rate estimates up, and down. that flights departing the U.S. tend to have some racial homogeneity flights departing for Japan have more individuals from East Asian countries than do flights departing to Nigeria, and more than would be expected by random selection. Another reason is that face recognition accuracy will be worse for homogenous galleries because false positives will be more common. Our practice of building homogenous galleries biases the test toward higher error.

4. The 12 regions are: Europe, W. Africa, E. Africa, N. Africa, Middle East, S. Asia, E. Asia, Oceania, N. America, C. America, Caribbean and S. America. We assign individuals to a region based on the issuers of their travel document. Occasionally some travelers will travel on a different countrys travel document; in such cases we assume their region to be that of the gallery ENTRY image.

5. We form 567 galleries, with one image per person. We form another set of 567 galleries with variable numbers of images per person. The number of galleries we can form per region varies because we have more images from some regions than others.

6. We search each gallery using a single probe-set containing 127 258 EXIT images of 123 075 people. By visual inspection it is evident that the images are collected using different cameras in different locations. For a given gallery only small proportion of the searches will have an enrolled mate in the departure gallery, at most 420 of 123 075 people. These mated pairs afford estimates of false negative identification rate. The remaining images, from persons of all regions of the world, form a non-mated search set used for estimating false positive identification rates.

[10] Aircraft configuration makes a difference, so that while the A380 is capable of carrying 560 economy class passengers it is atypical for that to occur for aircraft departing the United States.

[11] Not all, as the U.S. PCA stated, their TVS "does not enroll recent crossing images of U.S. travelers into the gallery, but does enroll recent crossing images of foreign nationals into the gallery."

[12] The single-image enrollment will be more pertinent to processing of citizens of a country for whom, often, only one photo exists in the gallery. The multiple-image enrollments yield better accuracy, and are pertinent to foreign travelers.

7. We run multiple algorithms, in some cases more than one from each developer. These were submitted to the one-to-many identification track of the FRVT between May 2018 and the present. The list of algorithms includes the NEC-3 algorithm that was broadly the most accurate through November 2018 as reported in NIST Interagency Report 8271 [1], but which has been eclipsed in accuracy by newer algorithms submitted since.

8. We compute 10 thresholds for each algorithm corresponding respectively to the 10 false positive identification error rates: 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1. We get the threshold value by looking at the highest non-mate score produced when running all non-mate searches against all galleries. Given, say, 126838 non-mate searches into each of 567 galleries, the threshold for FPIR = 0.0003 is taken to be the 126838 x 567 x 0.0003 = 21575-th highest observed rank-one comparison score.

3.2 Results

3.2.1 Attainable accuracy with single entry image

Figure 2 shows accuracy for two algorithms submitted to NIST 28 months apart. These are the NEC-3 algorithm submitted to NIST in November 2018, and the Visionlabs-10 algorithm from February 2021. The gallery size is N = 420 subjects, each person enrolled with exactly one ENTRY image. The vertical axis is a count of the individuals who are not biometrically authenticated during boarding. The horizontal axis shows the region of the enrolled population. The dots correspond of one departing flight. The dots are jittered horizontally around the region label, and vertically around the integer value, to avoid over-plotting and show the distribution.

The notable observations from the graphs are:

1. The number of false negative recognition errors is spread between zero and 16, with the most common value being 6. These errors would need to be resolved via a second attempt at biometrics, or via an airline-defined biographic process.

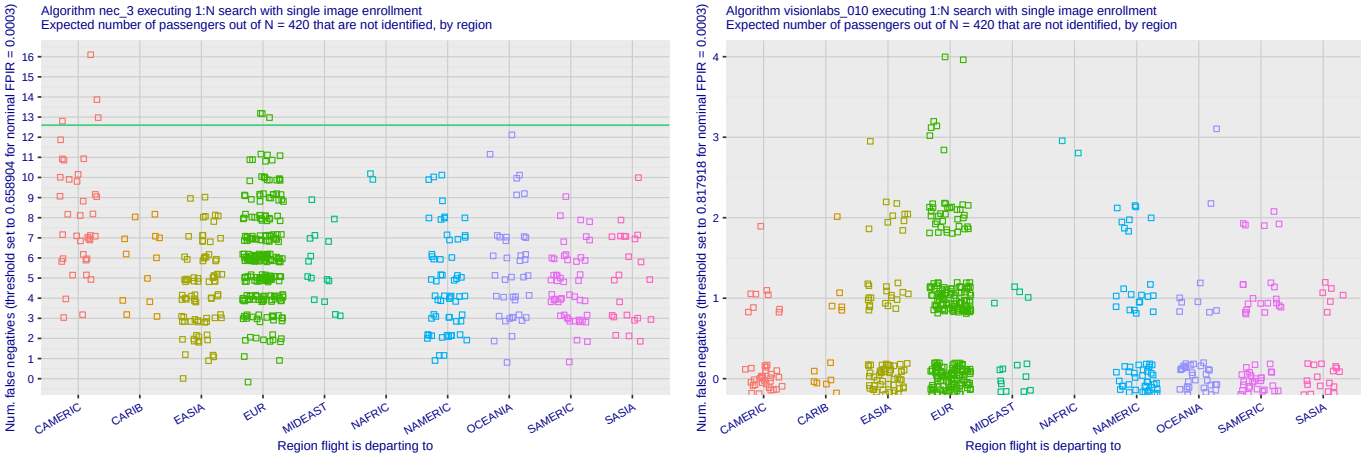
2. The distributions across regions are similar. The Central American flights give modestly higher FNIR, but this may simply be the result of chance. To the extent that some of the regions here are proxies for race, the results comport with those published in NIST Interagency Report 8280 [2] showing little dependence of false negative rates on race. Any false negative demographic differentials should be corrected for:

   (a) Ageing: It is possible that different travelers from certain regions travel less frequently such that the gallery photos are older time lapse affects appearance and accuracy.

   (b) Age: It is possible that absolute age affects accuracy. For example, although not the subject of the simulation here, flights into Orlando are disproportionately populated with children[13] whose lower height can affect head pitch angle and accuracy.

3. We report the rate of false negatives (FNIR) in a subsequent figure but note here that a count of 13 (i.e. 0.03 * 420) corresponds to a 3% failure rate. On that basis, the overall error rate is below 3% corresponding to better than the 97% verification rate required in 2007 legislation.

[13] In visits to observe EXIT boarding processes June 2019, the author observed children, without instruction, standing on tip-toes in order to present their face to the camera mounted above five feet. This was sometimes effective.



(a) NEC-3 (b) VISIONLABS-10

Figure 2: Count of false negatives on simulated flights by region using the NEC-3 algorithm from Nov. 2018 and the Visionlabs-10 version from February 2021. The gallery is populated with one ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The false negative identification rate for a flight can be stated by dividing the number of false negatives by the number of passengers, 420.

Figure 2(b) shows accuracy for a recent algorithm that is among the most accurate submissions to the one-to-many track of FRVT. The number of errors now is much lower, ranging from 0 to 4, with most common value being 0. A false negative count of zero corresponds to correct recognition of all passengers.

Note that the most accurate algorithms have been submitted to NIST recently, in early 2021, showing accuracy gains are still being realized by developer innovation. Several algorithms, including the VisionLabs-10 algorithm used in Figure 2(b), are more accurate than leading algorithms submitted to NIST in 2018 - see the ongoing FRVT webpage for names, dates, and more general accuracy results. The implication is that a PCA will realize accuracy gains if its technology refresh process is active and frequent.

Figure 3 shows the same figure for the most accurate algorithms tabulated appearing in Table 2. We note the following:

1. Figure 3 includes, in blue text, values for FNIR, the estimated proportion of passengers who will not be able to board with a single probe capture. The values are well near 0.1% for the most accurate algorithms, and often above 1% for the less accurate ones.

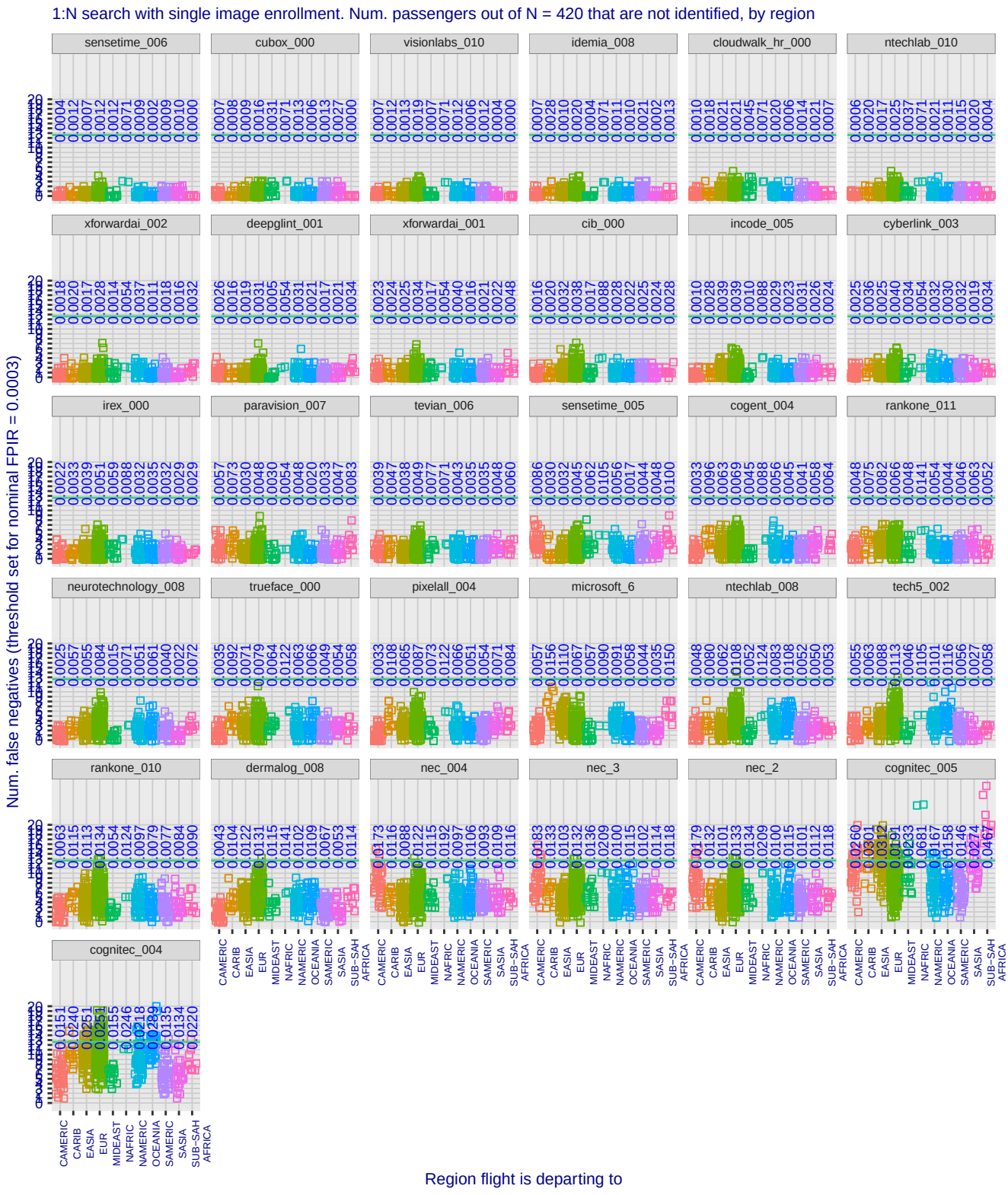


Figure 3: Count of false negatives on simulated flights by region. Each point corresponds to one flight with a gallery populated with one ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The blue text gives FNIR. The panels are arranged left-to-right, top-to-bottom in order of mean false negative count. The horizontal green line corresponds to the 3% false negative goal implied by legislation in the U.S.

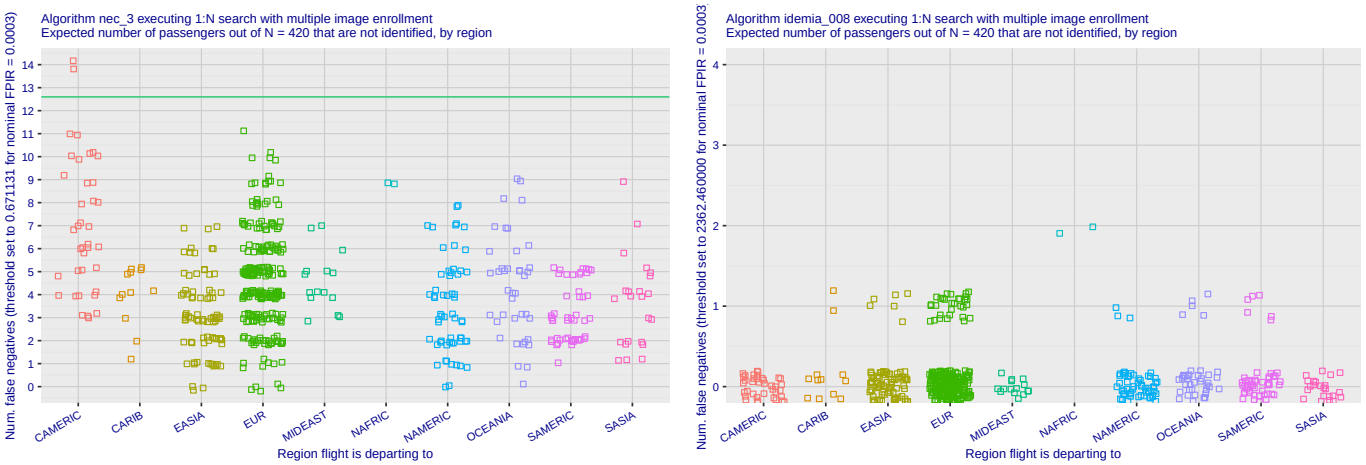
PCA = PASSPORT CONTROL AGENCY
TVS = TRAVELER VERIFICATION SERVICE
FNIR(N, R, T) = FALSE NEG. ID RATE
FPIR(N, T) = FALSE POS. ID RATE
N = NUM. ENROLLED SUBJECTS
T = THRESHOLD
$T = 0 \rightarrow$ Investigation
$T > 0 \rightarrow$ Identification

3.2.2 Attainable accuracy with multiple entry images

Figure 4 shows the accuracy results for two algorithms for a gallery of size N = 420 subjects each now enrolled with *multiple* ENTRY images. The algorithms were submitted 29 months apart, in November 2018 (NEC-3) and March 2021 (Idemia-8). From the two figures we note the following:

1. The use of multiple enrollment images reduces the number of false negative recognition errors modestly for NEC-3 (2018). It produces around 4 errors on average instead of 6 with a single image. The worst case count is reduced from 16 to 14.

2. With Idemia-8 (2021) the effect of enrolling more images is a more substantial reduction in false negative outcomes such that a large majority of flights will see all passengers board without any errors. The worst case count of error is reduced from 4 to 2.



(a) NEC-3 (b) IDEMIA-8

Figure 4: Count of false negatives on simulated flights by region using the November 2018 NEC-3 and March 2019 Idemia-8 algorithms. Each point corresponds to one flight the gallery for which is populated with multiple ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The horizontal green line corresponds to the 3% false negative goal implied by legislation in the U.S.

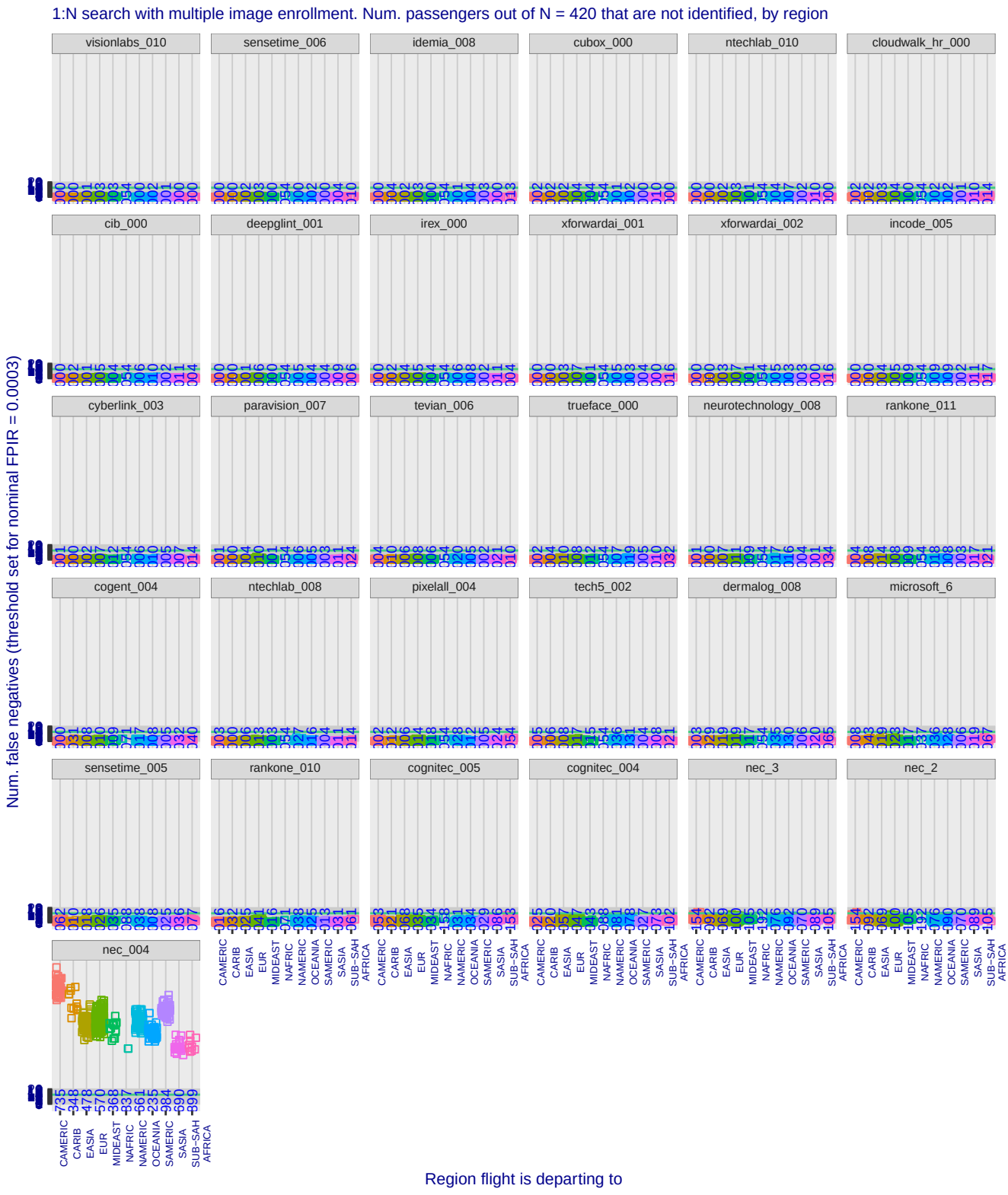


Figure 5: Comparison of false negative identification rates between number of images enrolled (one per person, vs. several) and between algorithms. The algorithms were submitted to between June 2018 and April 2021. Each point corresponds to one flight to the identified region the gallery for which is populated with ENTRY images from each of N = 420 individuals. The blue text is a false negative identification rate (FNIR), often below 1%. The orange text is the number of simulated flights, out of 567, for which the number of false negative errors is zero. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts.

The failure of NEC-3 to exploit multiple images may stem from how we provided images to the gallery. In FRVT we typically provide all images of an individual to the algorithm in one call to the template generation function the algorithm consumes multiple images and has the opportunity to select or fuse images as it sees fit. However, that is atypical operationally: images are provided to the algorithm serially such that multiple images of the same person result in separate enrolled templates - that is the model followed here[14] even though it denies the algorithms an explicit fusion opportunity. Figure 5 shows analogous results for nine of the more accurate algorithms evaluated in FRVT through November 2011. The panels are included in order of mean overall number of false negatives. Notably:

1. For the majority of flights, the most accurate algorithms correctly identify every passenger, and only ever fail to on up to 2 out 420 people.
2. On this metric there are multiple algorithms affording lower false negative identification error rates than does NEC-3. This is an existence proof of better accuracy that suggests an PCA will benefit from monitoring of test results and regular technology refresh. A PCA would need to factor other variables into procurement from a new developer including performance aspects (speed, scalability to large galleries, and demographic equitability) and contractual factors like capital and transaction costs, including those of integration.

3.2.3 False negative vs. false positive tradeoff

The results for each algorithm thus far have been stated at a single threshold. If we had set this threshold to a higher value the false negative rates would also have been higher, but with the advantage of lower false positive rates. Conversely, if the threshold had been low, false negatives would be better and false positives could occur more easily. The threshold is conventionally set to achieve a low enough probability that an impostor could match an enrolled identity thereby meeting some planned security objective. In one-to-many applications, an impostor only needs to match any enrolled identity to gain access he has N opportunities. This generally necessitates higher thresholds than one-to-one verification where the impostor claims one particular identity.

[14] See Figure 8 and section 3.2 in the FRVT 1:N report, NIST Interagency Report 8271 [1] for details on multi-image enrollment and metrics. See the FRVT page for newer algorithm results.

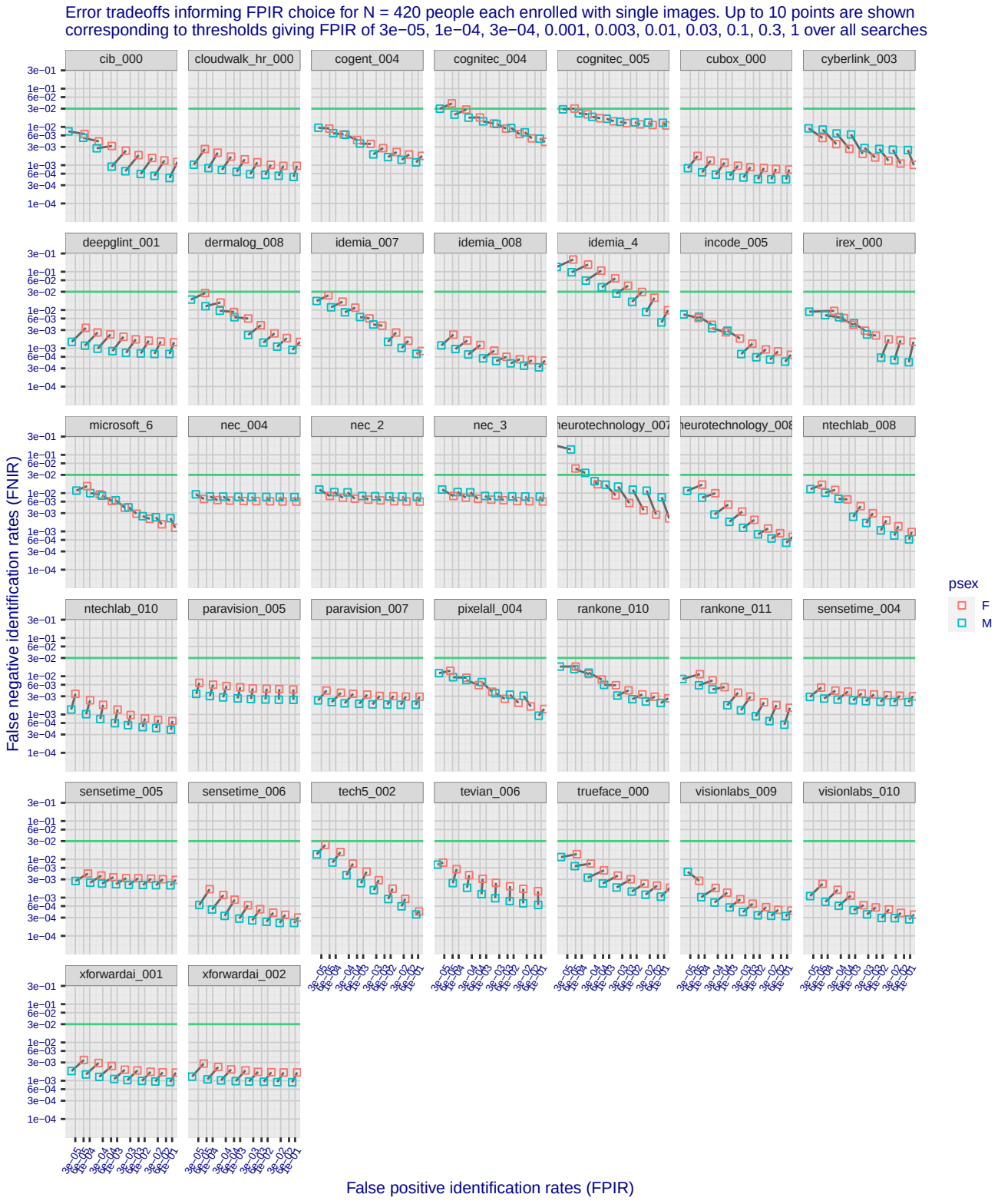


Figure 6: Error tradeoff for 26 algorithms executing 1:N searches with N = 420 people enrolled with a single image. The 10 point-pairs correspond to 10 possible thresholds for each algorithm. The red and blue boxes correspond to female and male travelers; their relative displacement indicates generally higher false positives and false negative rates in women. Smaller displacement indicates smaller (better) demographic differential (see NEC and Paravision-5, for example). The horizontal green line corresponds to 3% false negative goal implied by legislation in the U.S.

What should the false positive identification rate be?

This question is about policy. As discussed in in the introduction a TVS can serve double duty as an aircraft access control system and as a visa-holder EXIT status facilitation system. The discussion centers on what the system is trying to prevent stowaways, evasion of traveler-to-bag matching, or faking someones departure.

One factor is the prior probability that someone would try to board a flight at all. Its likely quite common and not necessarily nefarious - the author has accidentally gone to the wrong gate on more than occasion. For pure facilitation a low threshold could be used, but in its access control role that would allow any traveler to board, and potentially get free passage. A conventional value for access control is for false positive rate of 1 in 10 000. Lower values can used but impostors will switch to active attack techniques to achieve a false positive. One factor is variability in false positive rates with demographics: many algorithms can give 100 times more false positives on elderly, female people from certain countries.

Figure 6 shows the error rate tradeoff by plotting false negative identification rates against false positive identification rates at ten operating thresholds spread over four decades of FPIR, from 1 in 33 333 to nearly 1 in 3. Instead of showing the full curves, the ten-point pairs expose the increase in FNIR at low FPIR but also show the difference in error rates for men and women.

We note the following points:

1. Some algorithms give generally lower FNIR across the range of FPIR. This is simply a re-iteration that accuracy varies markedly.
2. Some algorithms give a flat error tradeoff characteristic. This is most evident for the idemia-8, deepglint-1, nec-3, paravision-5 and sensetime-4 algorithms. This is an attractive property of any biometric system because it allows very low false positive identification rates to be attained without intolerable increases in false negative identification rates. This becomes important later when we increase the enrolled population size by a factor of 100.
3. Most algorithms give FNIR below 0.03 (the green line in the plots) for a wide range of FPIR, thereby meeting the legislative mandate to be able to verify the EXIT of 97% of (in-scope) travelers.

Comparing Figure 7 with Figure 6 shows that across the four-decade range of FPIR, the FNIR values are reduced by using multiple enrollment images. The single-image enrollment represents "worst-case" of having just one prior encounter. The multi-image case is more typical.

Note that this analysis doesnt answer the technical question of whether enrolling multiple images per subject increases FPIR versus using just single image. The reason is that the thresholds for multiple enrollments are generally higher than for singles. There are exceptions Idemia-7 for example. The question is important in situations where some travelers might have dozens of enrollment images and the algorithm response could be to attract false matches i.e. to make such enrollees lambs[15].

[15]The term lamb, a category defined in "The Biometric Zoo", refers to an enrollee who attracts more than average number of false matches.

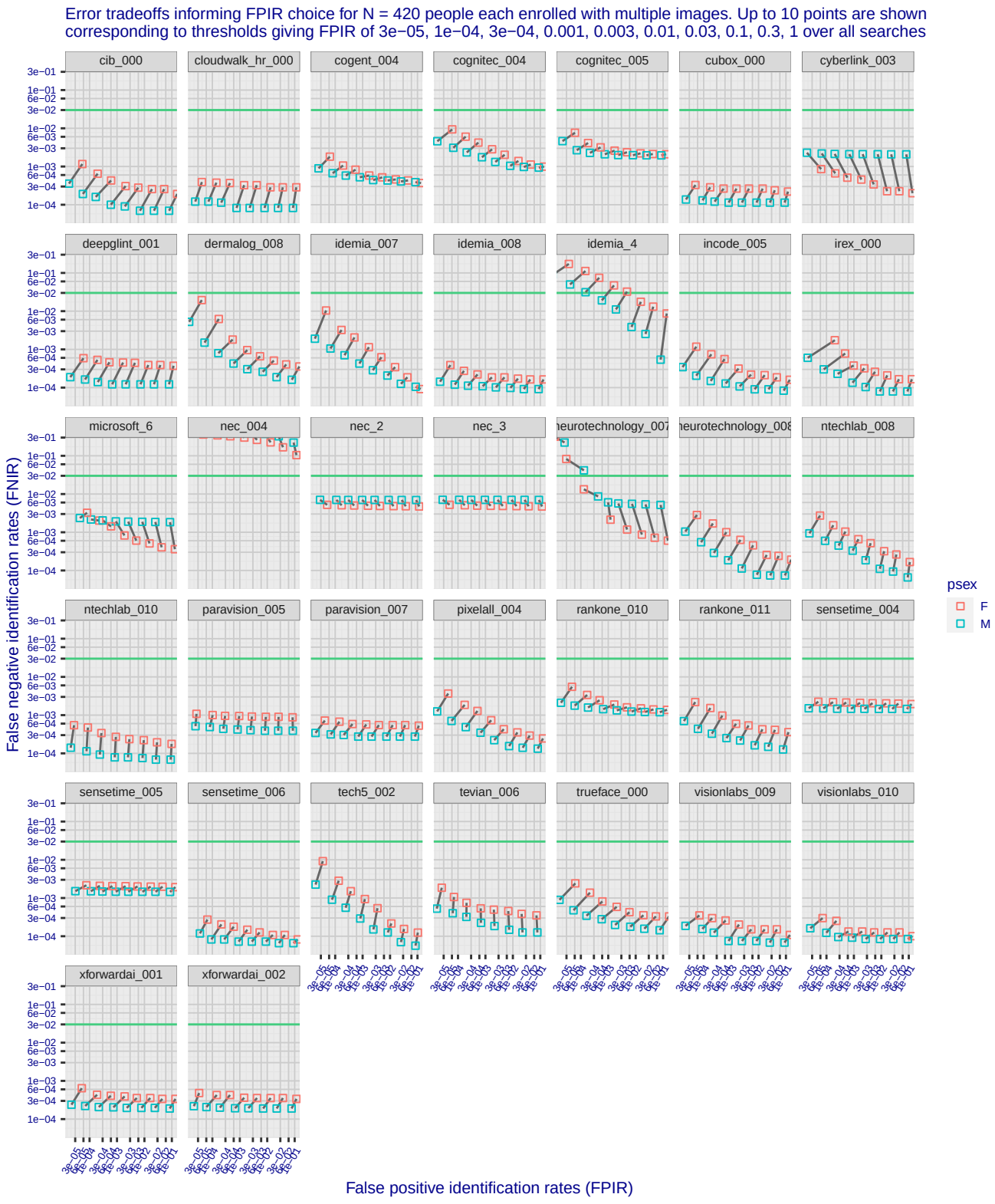


Figure 7: Error tradeoff for 26 algorithms executing 1:N searches with N = 420 people each enrolled with multiple images. The 10 point-pairs correspond to 10 possible thresholds for each algorithm. The red and blue boxes correspond to female and male travelers; their relative displacement indicates generally higher false positives and false negative rates in women. Smaller displacement indicates smaller (better) demographic differential (see NEC and Paravision-5, for example). The horizontal green line corresponds to 3% false negative goal implied by legislation in the U.S.

3.2.4 Demographics: Differentials by sex

Vertical displacement of point pairs in Figure 6 and Figure 7 reveal broadly higher False Negative Identification Rates in women than in men. This is consistent with NIST IR 8280 using other kinds of images. The cause of this is not known. Note some algorithms, including NEC-3, Microsoft-6 and Neurotechnology-7, give the opposite behavior or fairly equitable rates.

The horizontal displacement in the figures show that all algorithms give a factor of 2 or 3 times higher false positive identification rates in women; this means that women will be mismatched against a wrong identity somewhat more often than men. This will be rare but over enough flights it will disadvantage more women than men. Algorithms from Microsoft, NEC, and Cognitec give notably smaller differentials.

Figure 8 summarizes false negative rates by sex: the difference often amounts to 1 additional false negative in women than men. Note that there are many flights with zero false negative flights for both sexes.

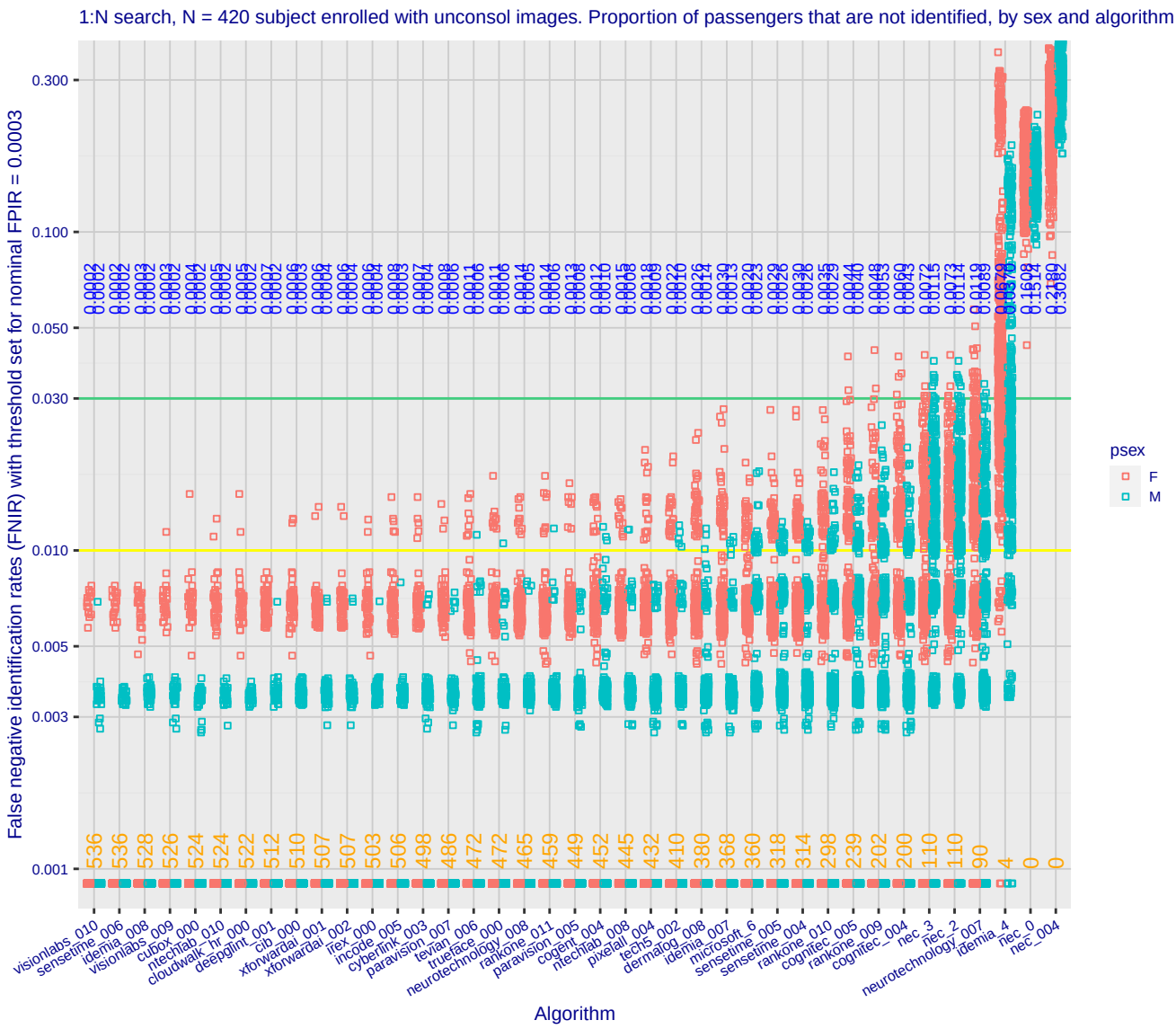


Figure 8: False negative identification rates by algorithm and sex. Each point corresponds to the boarding of N = 420 people on to one flight where each is enrolled with multiple images. The blue text is a FNIR value for that algorithm on that sex. The green line connotes a 3% FNIR (reflecting a legislative mandate). The yellow line is at 1% FNIR. The cluster of points at 0.0009 corresponds to zero errors (adjusted to plot on a log scale) - the orange text gives the number of simulated flights, out of 567, for which there are no false negative errors. The next cluster near 0.004 corresponds to 1 error out of around 210 males.

3.2.5 Demographics: False positve differentials by region

Figure 9 shows false positive identification rates by region and by sex for two algorithms from NEC and Canon. Appendix B gives analagous figures for all algorithms. We make the following comments:

1. **Magnitude:** The NEC-3 algorithms shows FPIR is quite insensitive to geography and sex with false positive identification rates estimates mostly clustered between $2x10^{-4}$ and $7x10^{-4}$ In constrast Canon's cib-000 algorithm gives FPIR estimates between $7x10^{-5}$ and $2x10^{-3}$. As noted in NIST Interagency Report 8280 the NEC-3 algorithm is taking steps to normalize false positive rates in one-to-many searches.

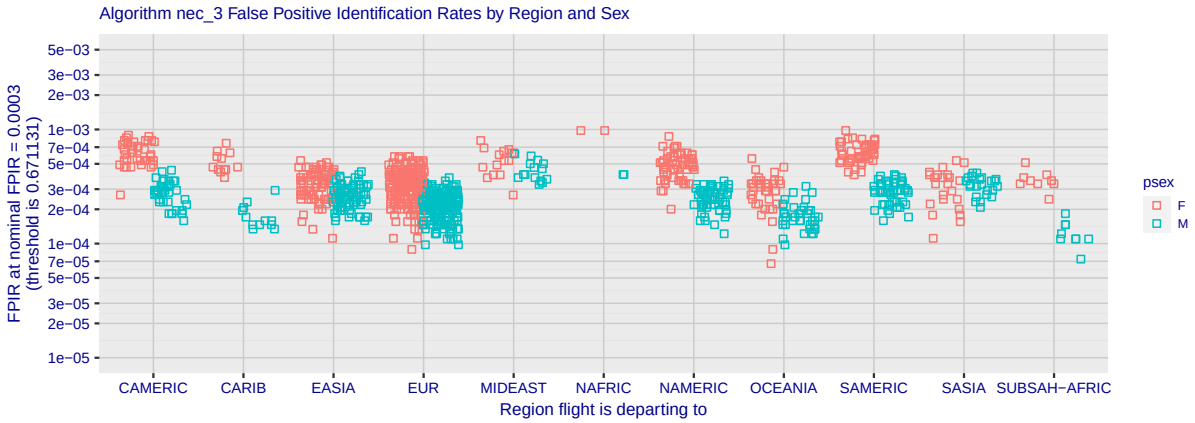
2. **Sex:** It is very common across algorithms for women to give higher FPIR than men. The NEC-3 algorithm gives

broadly the smallest differential in FPIR value cf. the y-axis in the next Figure.

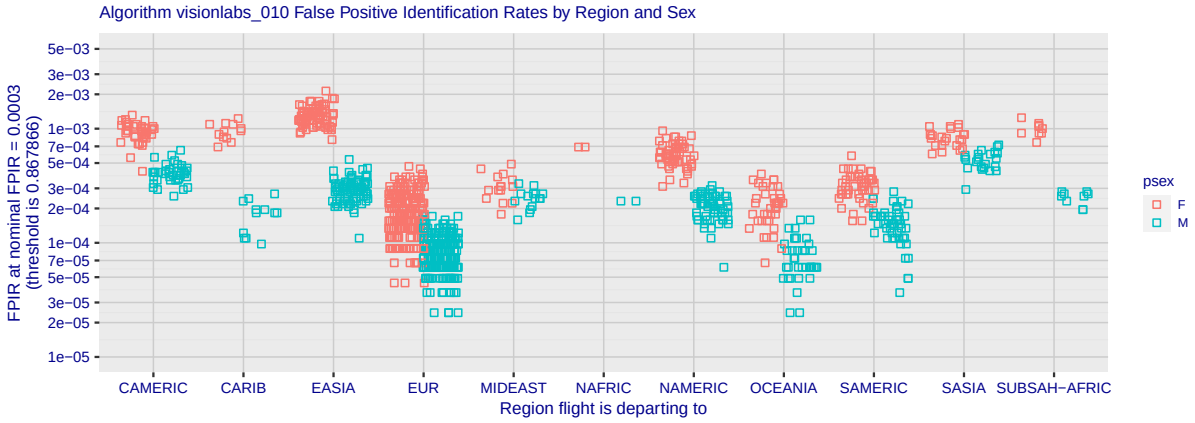
3. **Region:** False positive identification rates are commonly an order of magnitude higher in Asian women than in European men. For Canon's cib-0 algorithms there is a factor of 30 variation.

As always, with the observation of a demographic differential, the question is "what is the impact"? The overall target FPIR was 0.0003, achieved by setting an algorithm specific target. The worst upside departure from that is Canons cib-1 algorithm (see Appendix B) which gives FPIR for Asian women near 0.003. This FPIR, 1 in 333, is still low but implies between one and two false positives per flight boarding these would likely manifest as an in-gallery false match described in section 1.5. This may be an acceptable cost, but does constitute a disadvantage for Asian women attempting to record their departure from the United States.

The error tradeoff characteristics of figure 11 are, for some algorithms, quite flat implying that even lower false positive identification rates could be targetted (by increasing the threshold) without great adverse implications for false negative identification rates.



(a) NEC-3



(b) VisionLabs-10

Figure 9: For two algorithms, each point shows a false positive identification rate estimated by running c. 120 000 searches against that flight's gallery. Red and blue connote male and female enrollees. Analogous figures for all algorithms appear in Appendix B.

4 Scaling One-To-Many Authentication to Larger Populations

4.1 Motivation

Thus far we have discussed the use of 1:N face recognition for recording the exit of travelers while boarding an aircraft. A PCA can appropriately limit enrollment in FR galleries to just the population expected on the flight. This data minimization reduces mis-match possibilities. However, the travel industry has articulated a vision for paperless travel. In its simplest form, this starts when a traveler authenticates to an authoritative travel document (passport) using a one-to-one biometric verification of a live photo and then proceeds through an airports touchpoints such as the TSA line and airline lounges and aircraft boarding, without presenting a boarding pass. Instead, the traveler engages a camera which submits a photo as a query into a database of individuals expected and authorized to proceed. For example, such a system could be fielded at a security checkpoint. In such cases many more people would need to be enrolled into the face recognition engine than at a departure gate for example, all people expected in the airport during a time window extending from a few hours before their respective flights to the time of expected or actual departure. The number of individuals could readily extend into the tens of thousands, and more if airside locations would additionally recognize inbound passengers (e.g. buying in duty-free shops).

4.2 Background

The dependence of recognition accuracy on enrolled population size is well known. Qualitatively, as enrolled population grows any given search has a greater possibility of a false match. Such outcomes can occur for two types of traveler.

1. An illegitimate traveler someone who is not expected in the airport makes a presentation to the camera in attempt to pass the checkpoint. This succeeds if the traveler matches any enrolled identity with a comparison score above threshold.
2. A legitimate traveler someone who is expected at the touchpoint presents to the camera but matches an identity other than self. This may be inconsequential at a TSA line, but would be consequential in a hypothetical duty-free store application of this approach should the biometric result allow purchase without further authentication.

The rate at which false positives occur is the false positive identification rate (FPIR). In a biometric test, FPIR is estimated by conducting non-mated searches into an enrolled population. FPIR is stated as the number of searches resulting in a false positive divided by the number of non-mated searches. How does FPIR scale with the number of enrolled identities? There are two classes of face search algorithms: Class A is those that implement a 1:N search as N 1:1 comparisons followed by a sort operation, and Class B is comprised of everything else including those that implement some more complex search strategy.

- **Class A** algorithms are expected to give a FPIR increases with the number of enrolled identities. It may increase further also if those identities are enrolled with several images each. Given a system in which N people are enrolled, with one image each, a standard binomial model gives

$$\text{FPIR}(N, T) = 1 - (1 - \text{FMR}(T))^N \tag{3}$$

  where the system owner sets the threshold T, and has an estimate of FMR(T), the false match rate in purely one-to-one comparisons. For small FMR, this approximates to

$$\text{FPIR}(N, T) = N\text{FMR}(T) \tag{4}$$

implying that the one-to-many false positive hazard grows linearly with N.

- **Class B** one-to-many algorithms are those that do not implement 1:N search using N 1:1 comparisons - these can include fast-search algorithms (using trees, indexes etc) and those that normalize scores across some or all of the gallery entries. These algorithms may not exhibit the (near) linear dependence of equations 3 and 4. This can occur for other reasons also. Some algorithms adjust comparison scores to the database size such that FPIR becomes approximately independent of N. The NEC-3 and Idemia algorithms exhibit such behavior (see FRVT Part 2 and its report cards). This relieves the system owner of the need to configure thresholds for the given population size. A system owner might consult vendor documentation, or consult NISTs FRVT Part 2 report which documents the dependence of FPIR on N and T.

4.3 Simulation of Large-N Accuracy

4.3.1 Experimental design

We repeated the EXIT simulation given previously but instead of enrolling N = 420 individuals with one ENTRY image, we mixed in a further 41580 such images from a disjoint population selected without regard to demographics. The result is a set of 567 galleries, each with N = 42 000 individuals. This population size is somewhat larger relative to the number of international passengers appearing daily in large U.S. airports in 2019.

4.3.2 Results

Figure 10 shows the number of false negatives expected when using the algorithm named on the horizontal axis to search three different kinds of galleries. The first enrolls 420 people with a single PCA ENTRY image; the second enrolls those same people with all prior PCA ENTRY encounters; the last enrolls 42 000 people with a single image. The kind of gallery is encoded by the shape. The vertical position of each point is the mean (over 567 regional galleries) of the number of false negatives when 420 test subjects are searched. The color of each point encodes the fraction of all 567 trials that give three or fewer false negatives.

We make the following observations:

1. The number of false negatives is higher with N = 42 000 than with N = 420, as expected.
2. Some algorithms nevertheless give only modest increases in the number of false negatives. In the most accurate case, the mean number of passengers being rejected would be below 1% (4/420), and more than 75% of flights (trials) would have three or fewer false negatives out of the 420 people making attempts.
3. Other algorithms give substantially higher false negative rates the graph shows FNIR approaching about 8% (34/420) for legacy algorithms.
4. Note that this analysis does not consider variance around the point estimates, nor sex or regional differences.

4.3.3 Discussion

Large enrolled populations require algorithms to be configured to operate at lower false match rates following Equation 4 an increase in N from 420 to 42 000 will necessitate a 100-fold FMR reduction to maintain constant FPIR. This puts a premium on algorithms that maintain relatively low FNIR at lower FPIR.

By inspecting Figure 11, for N = 42 000, all algorithms except some from Cloudwalk, Deepglint, Idemia, NEC, Paravision, Sensetime, Visionlabs and X-ForwardAI cannot maintain FNIR below 0.03 so, depending on FPIR, would not be meeting a legislative mandate for FNIR < 0.03 and TPIR > 0.97.

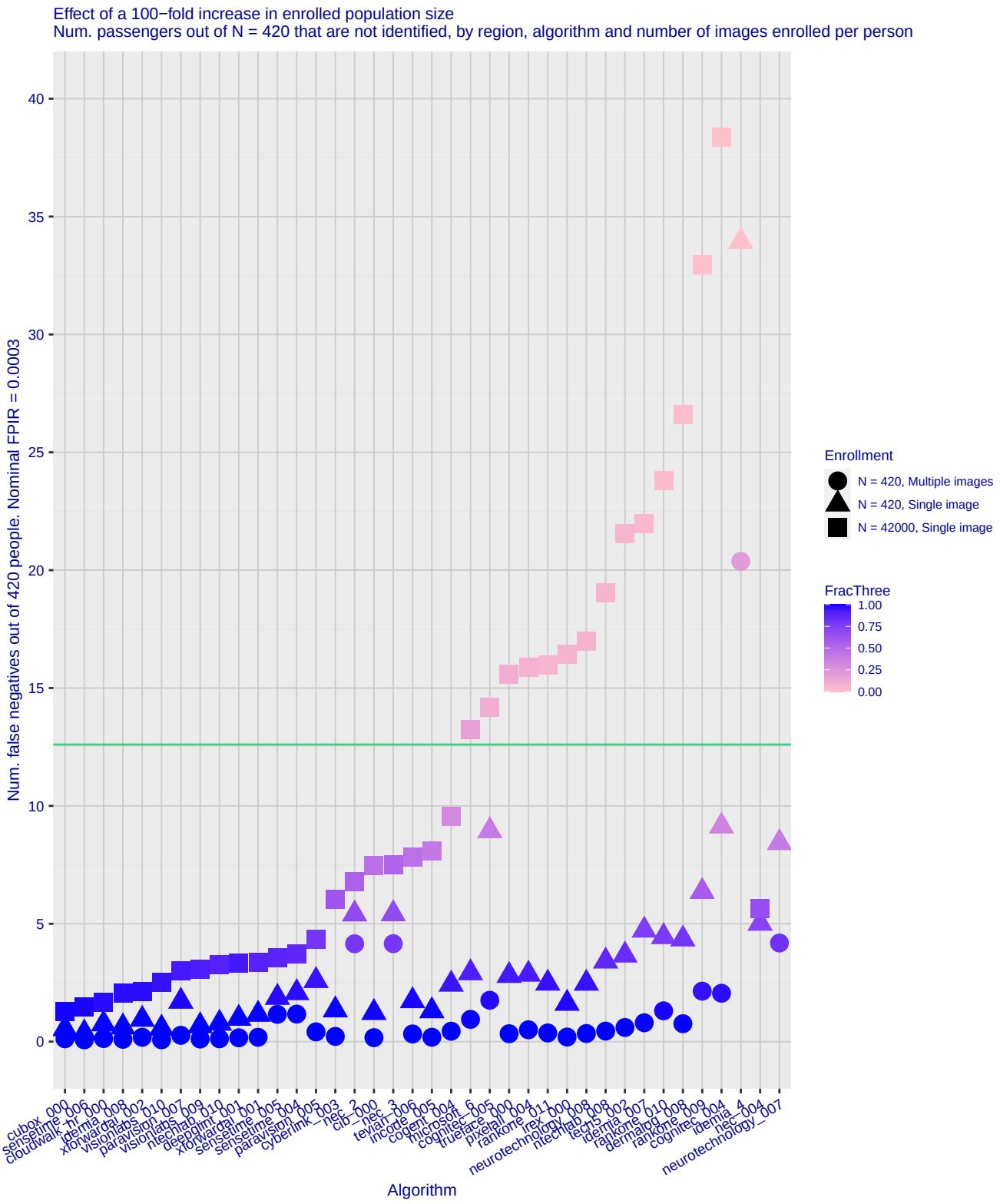


Figure 10: Comparing the number of false negatives expected when N = 420 people are searched against galleries containing imagery from enrolled populations of size of N = 420 and N = 42 000. The threshold is fixed in all cases to produce a false positive identification error rate of 1 in 3 333 (FPIR = 0.0003). The threshold value for each algorithm will usually be higher for the larger gallery to maintain the same false positive likelihood. The horizontal green line corresponds to a 3% false negative rate.

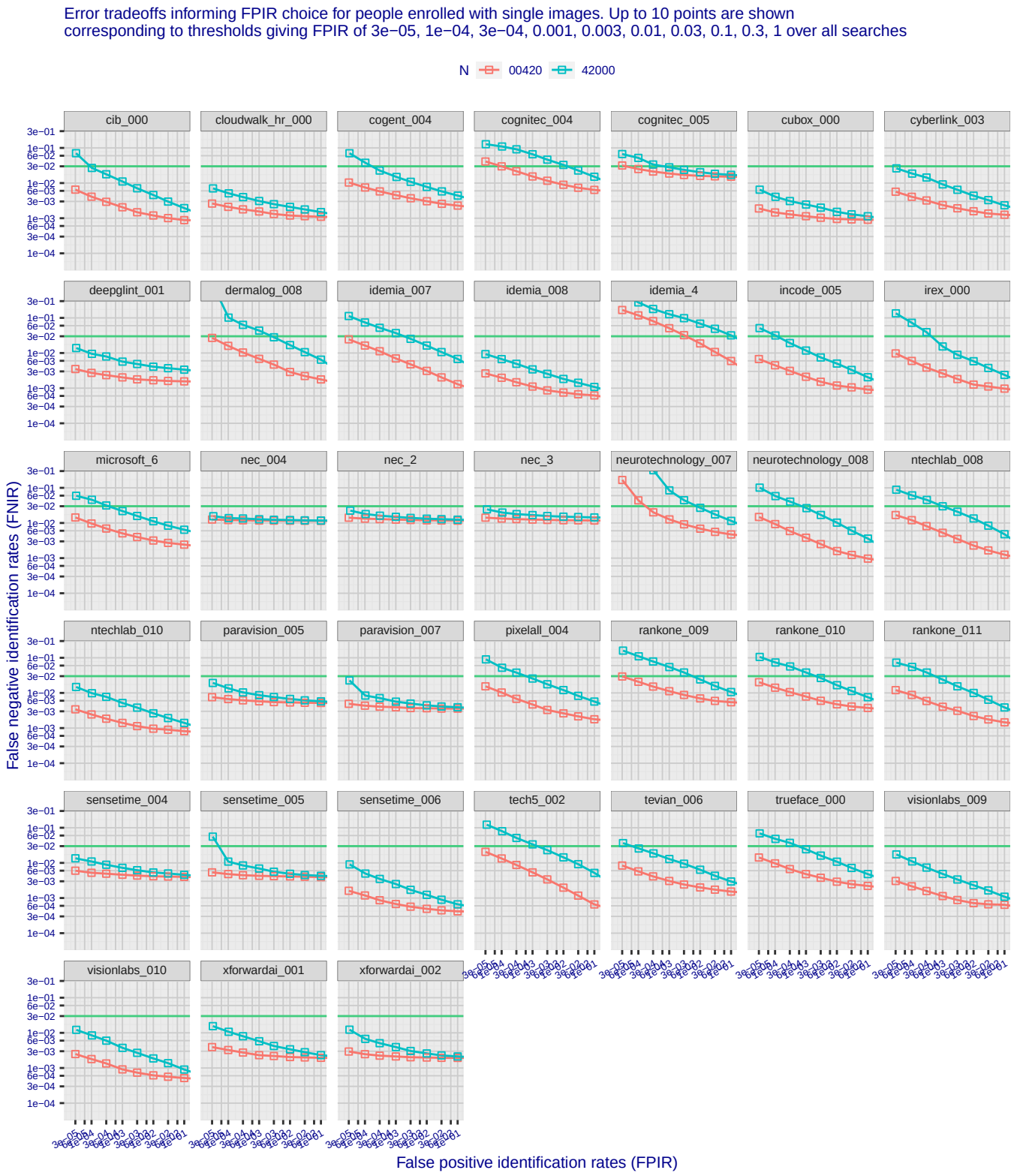


Figure 11: Error tradeoff characteristics for twelve algorithms conducting identical sets of searches into galleries of size N = 420 and N = 42 000. The horizontal line corresponds to a 3% false negative identification rate. The left side of each panel is relevant to the more "lights-out use of FR in positive access control and EXIT facilitation; the right side of each panel corresponds to high false positive identification rates for investigative uses of FR where humans review candidate lists. A flat profile confers the advantage of being able to run at lower FPIR without much elevation in FNIR.

5 Factors That Render Accuracy Estimates Approximate

The result in this report do not constitute an answer to the questions "how well does a particular TVS work", "does TVS satisfy a 97% legislative verification mandate" or "what is the accuracy of a PCA's EXIT solution". Why? Because the questions are different and because the tests we have reported here, while extensive, depart from the intended and desirable tests as follows. For each factor discussed, we note in blue the expected effect on accuracy.

1. **Airline re-direction of passengers.** During the EXIT pilot, airlines diverted some customers toward the legacy paper-based boarding process. This was particularly true when boarding was proceeding slowly or when cameras of the network to TVS were malfunctioning. This is not expected to bias accuracy in an offline test either way, but would lead to complication in using TVS logs to measure accuracy.

   The population so diverted was sometimes not random very tall or short travelers, and those with children, would be directed from the FR line. To the extent that this occurs, and to the extent that the NIST EXIT collection is not itself affected by this, our estimates of accuracy may be too high.

2. **Algorithms:** NIST does not have access to the actual algorithms deployed in TVS systems. Instead this study uses prototypes submitted to the one-to-many search track of the FRVT. These prototypes are identified using a name and a number. For example, "NEC-3" is from the NEC Corporation, and the three is simply a sequence number of algorithms sent to NIST. NIST is unable to confirm whether any prototype in FRVT has ever been deployed. Indeed a developer may make decisions on whether to productize a prototype on the basis of FRVT-derived technical information. In any case, a developer will maintain their own versioning designations. NIST is not provided with copies of operational algorithms.

3. **Active development**: Given persistent improvements in accuracy, as documented in FRVT, it is incumbent on end-users to instantiate a "technology-refresh" procedure so as to realize accuracy gains. Note that results in this report for 2018-era algorithms are likely out-of-date. Thus, for any given developer, it is likely that higher accuracy is available than is estimated here.

4. **Algorithm post-processing:** Accuracy will change if any software is used to post-process candidate lists produced by the algorithm. Conventionally the face recognition algorithm issues a candidate identity and a similarity score, which is compared to a system-wide threshold. If post-processing is used to re-score or re-rank then its effect on both false positive and false negative identification rates should be measured by comparing with that available from the raw candidate lists alone.

5. **Image data:** International travel has long been predicated on presentation of a passport. With e-Passports it is common for passport images to be retrieved and used for 1:1 verification of the the traveler. If on ENTRY those images are retained by the PCA , they can be used in downstream EXIT face recognition processes. The same applies to visa portraits collected as part of a visa application.

   (a) We did not have passport or passport-equivalant images for use in this study. These include visa images of various travelers and passport photos of the citizens persons. Instead we used airport arrivals hall photos with reduced quality. To the extent that a TVS makes use of high-quality passport and visa images, the accuracy values reported here are likely to be worse than for a system for which such images are available.

   (b) In this study we used an extract of a much larger corpus provided to NIST in May 2019. These images were anonymized and accompanied by limited metadata. The set included images labelled *exit*, and *ENTRY*. The former were collected in airport departures. A few of the exit images appear to have been collected from

persons in a vehicle, as could occur at a land-border. This factor will tend cause our accuracy estimates to differ from those of an operational TVS.

(c) Our exit images were collected in 2018 and the first four months of 2019. We assume that subsequent cameras, and their refined deployment by airlines, will yield improved images today compared to those used in this study, so we would expect improved accuracy over that noted here.

(d) Moreover, our exit images are not accompanied by camera make and model information, nor flight manifest information. It was therefore not possible for NIST to exactly reconstruct "a flight" instead we pooled all exit images as probes searched against each gallery. Our search set therefore pools exit images from quite different cameras and locations (airports). We are therefore unable to compare cameras and collection sites. From observations made during site visits, we note markedly different approaches to the quality-speed tradeoff. We expect therefore that our accuracy estimates have reduced variance compared to that from a TVS.

(e) If exit images are retained, even for a short period, they may be useful in offline "after-hours" accuracy estimation. For example, images from one flight could be used to make non-mated searches into a gallery of another flight, so as to estimate FPIR.

(f) Our galleries were constructed to hold people from one travel region as inferred from the nation that issued their travel document. This means that the galleries in this document will contain people who never flew together.

6. **Homogenous galleries.** Our practice of making galleries from people holding travel documents from countries in the same region of the world probably means that false positive rates are higher than if the galleries had been composed of a more mixed population. This practice would tend to depress our accuracy relative to those in TVS.

7. **Presence of images active attack.** It is possible that some of the captured images are from a presentation attack that went undetected, for example using a face mask. The occurrence of this is considered to be very small. Note that since we didnt have passport images, we do not expect the dataset to contain morphed images. While this is increasing possibility operationally, it can be averted by live, trusted capture of images as in a primary passport control lane.

Appendices

Appendix A Figures summarizing false negatives for each algorithm

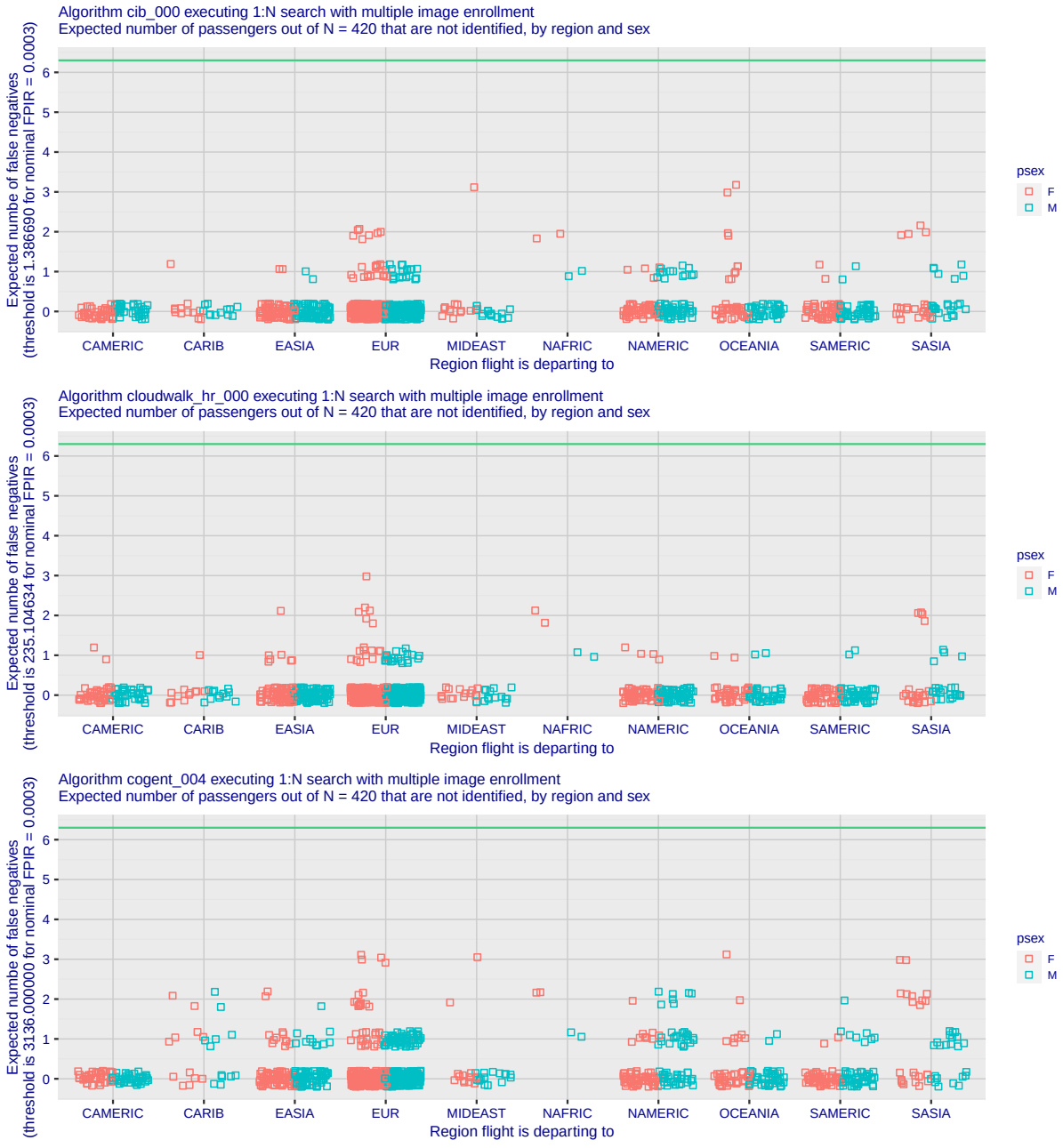


Figure 12: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

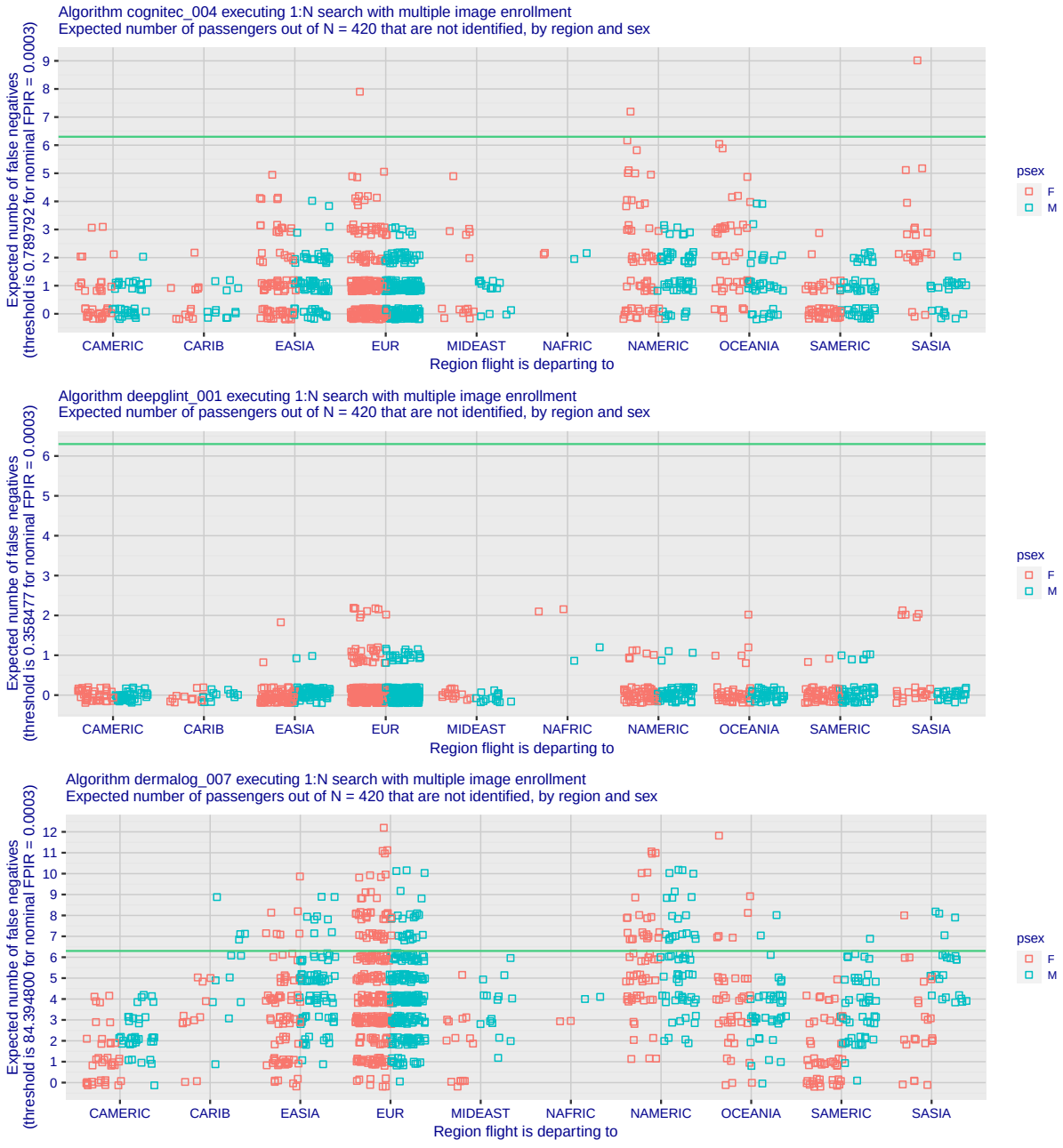


Figure 13: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

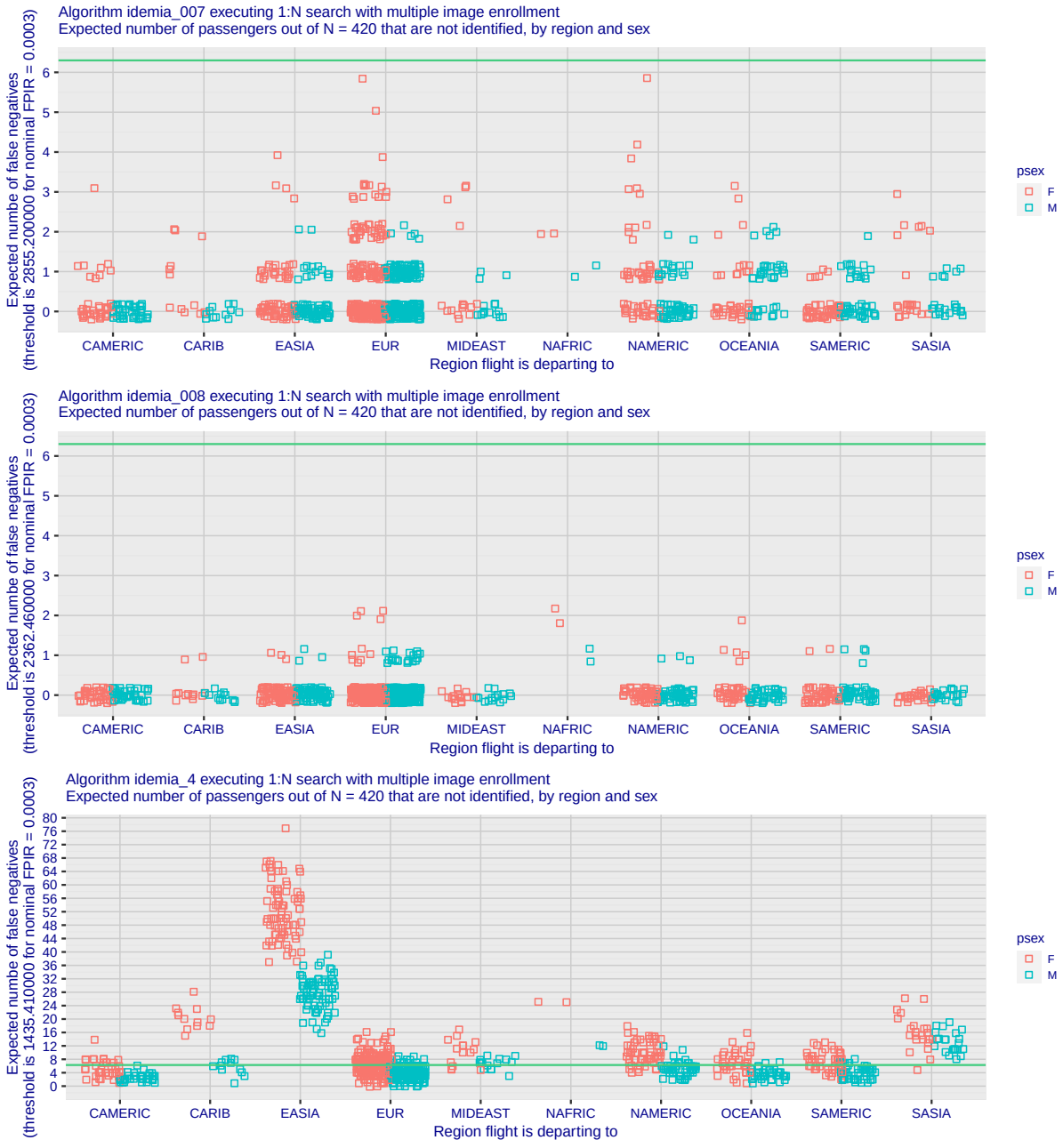


Figure 14: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

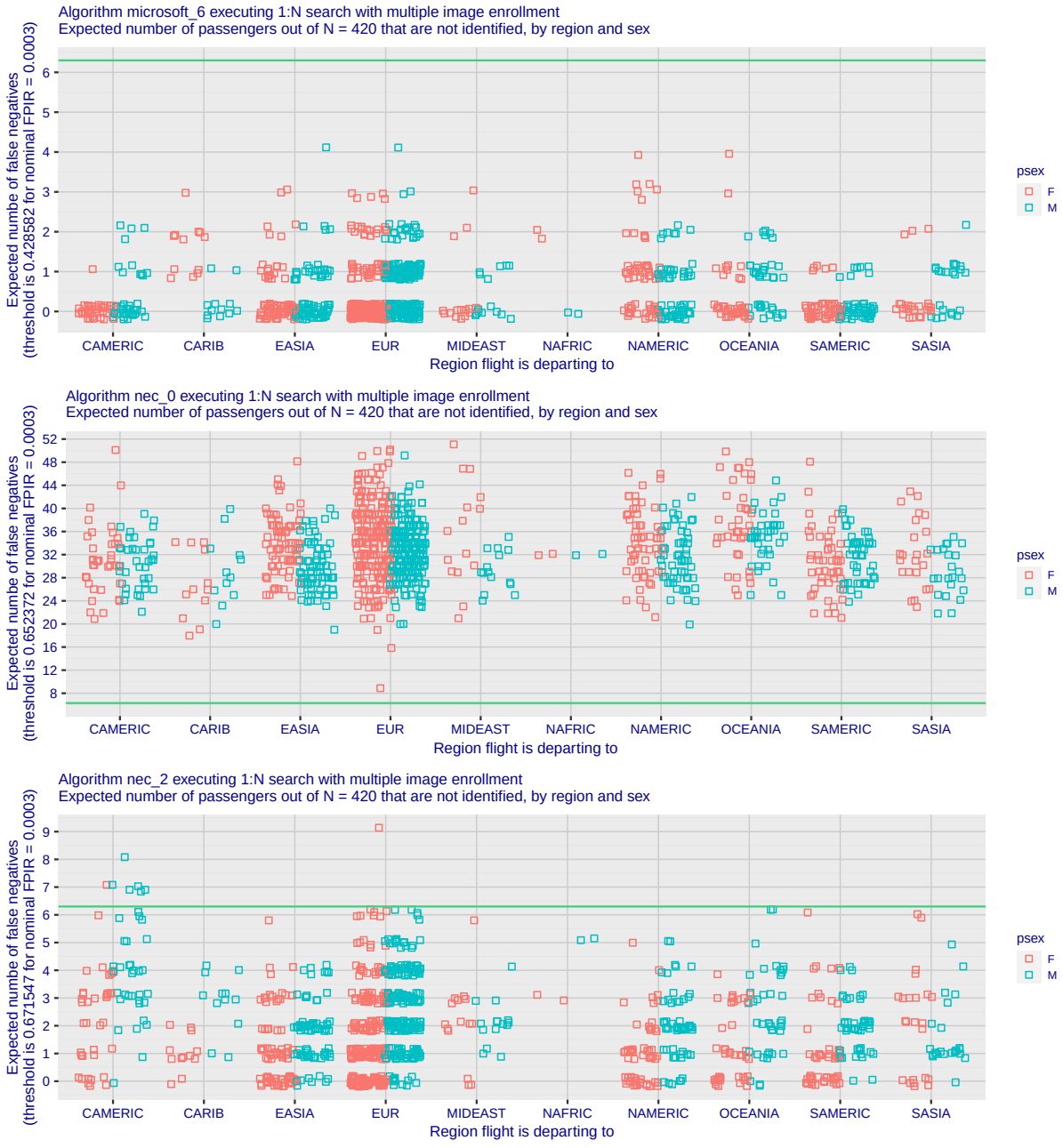


Figure 15: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

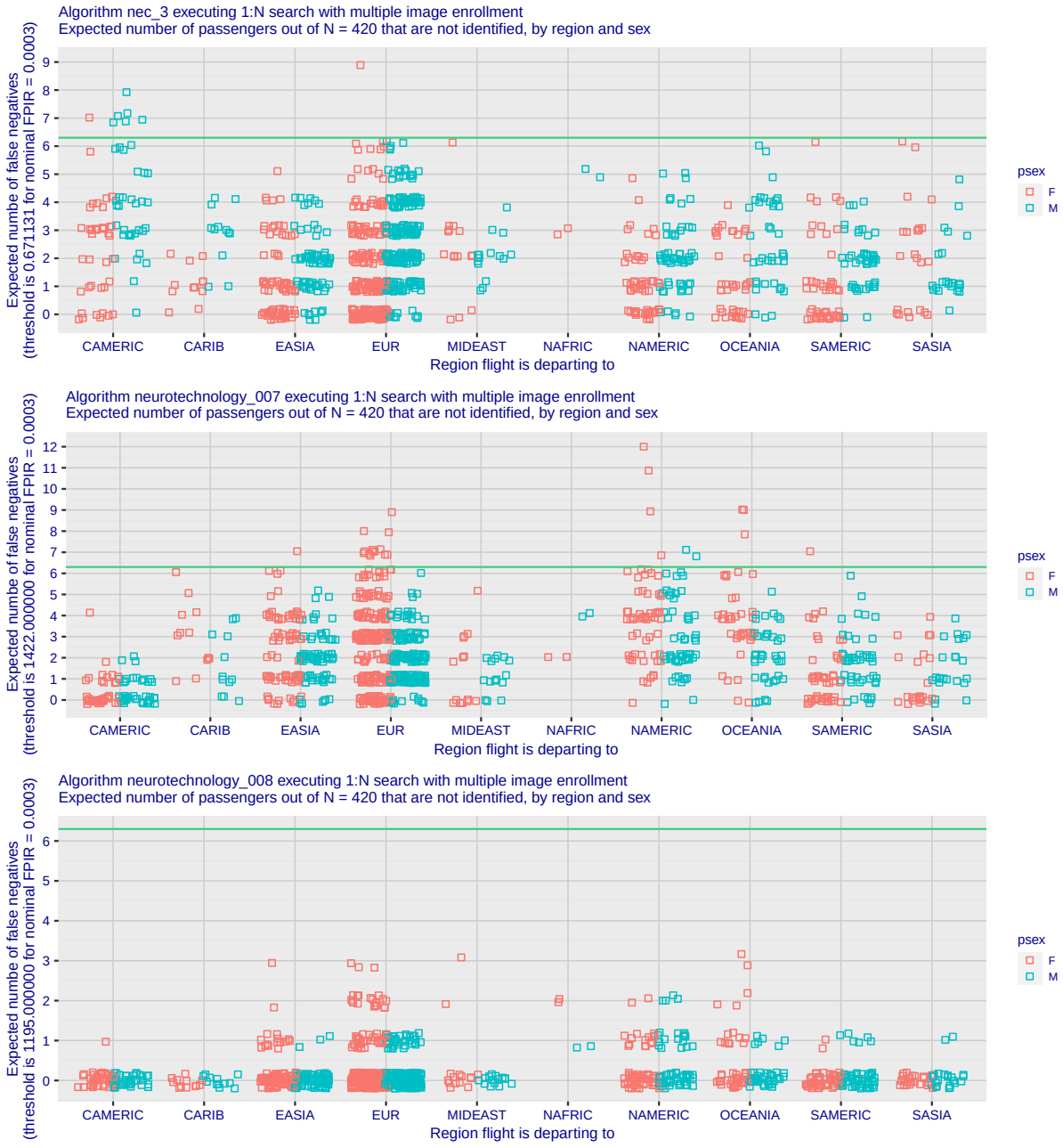


Figure 16: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

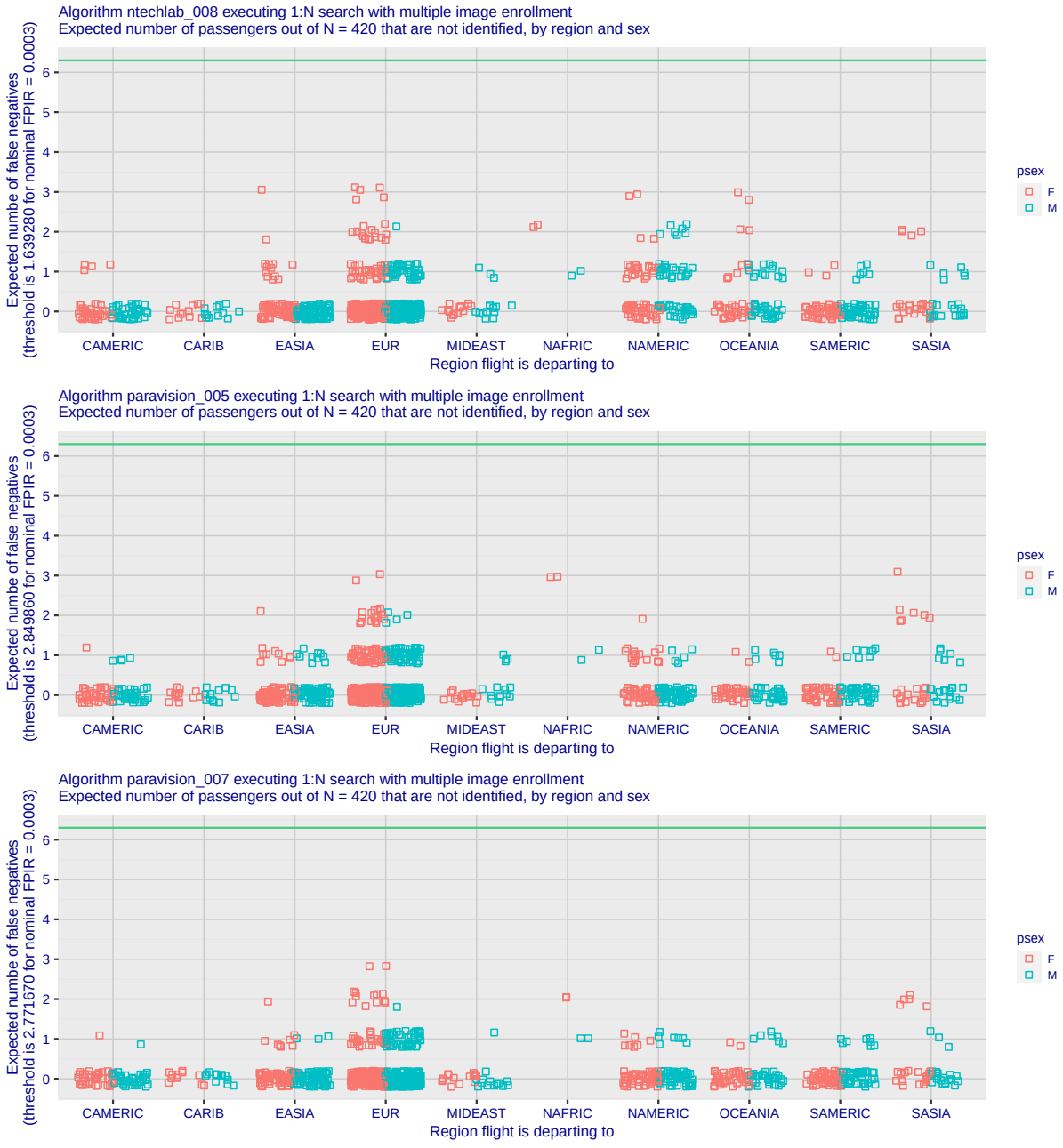


Figure 17: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

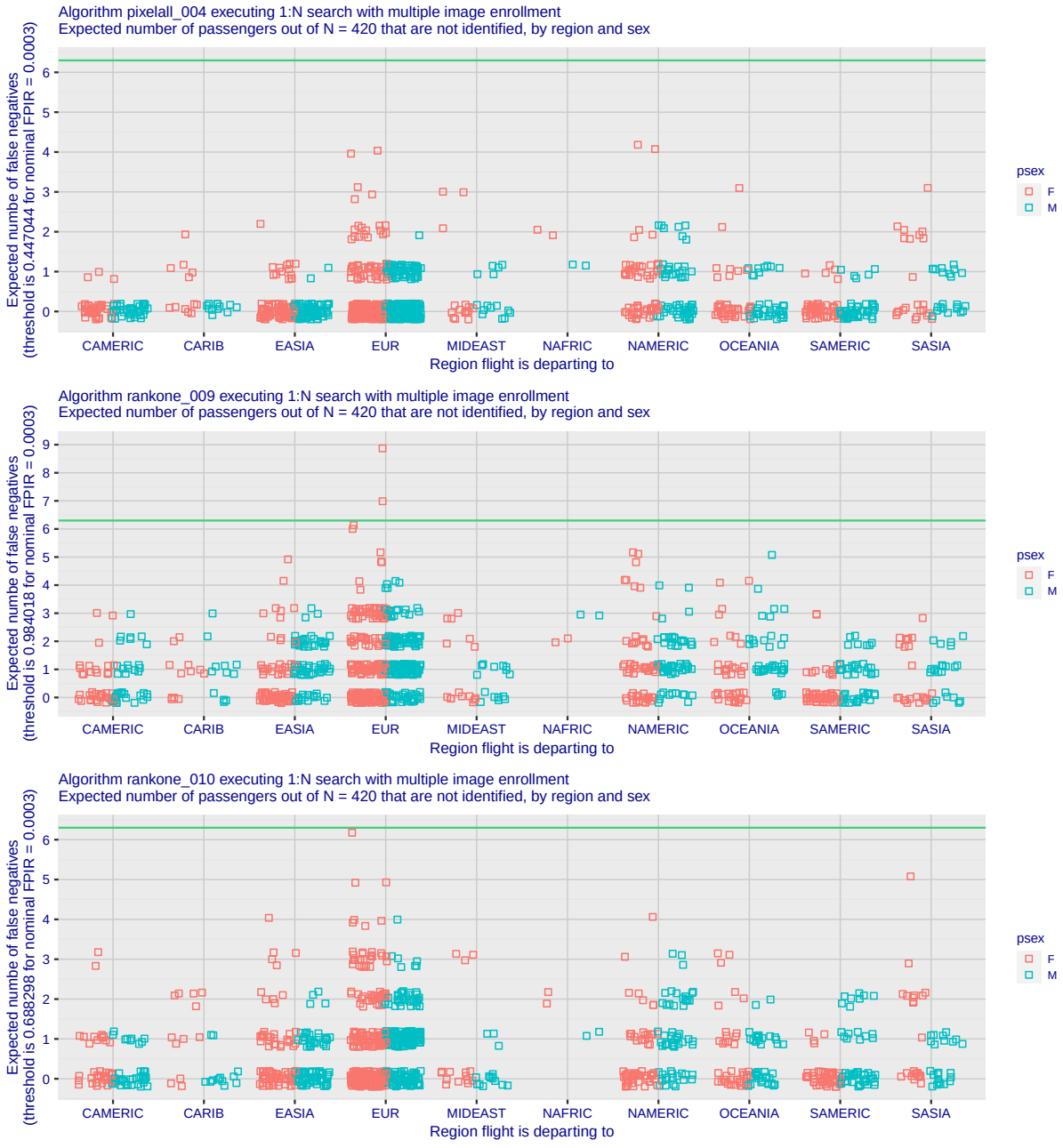


Figure 18: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

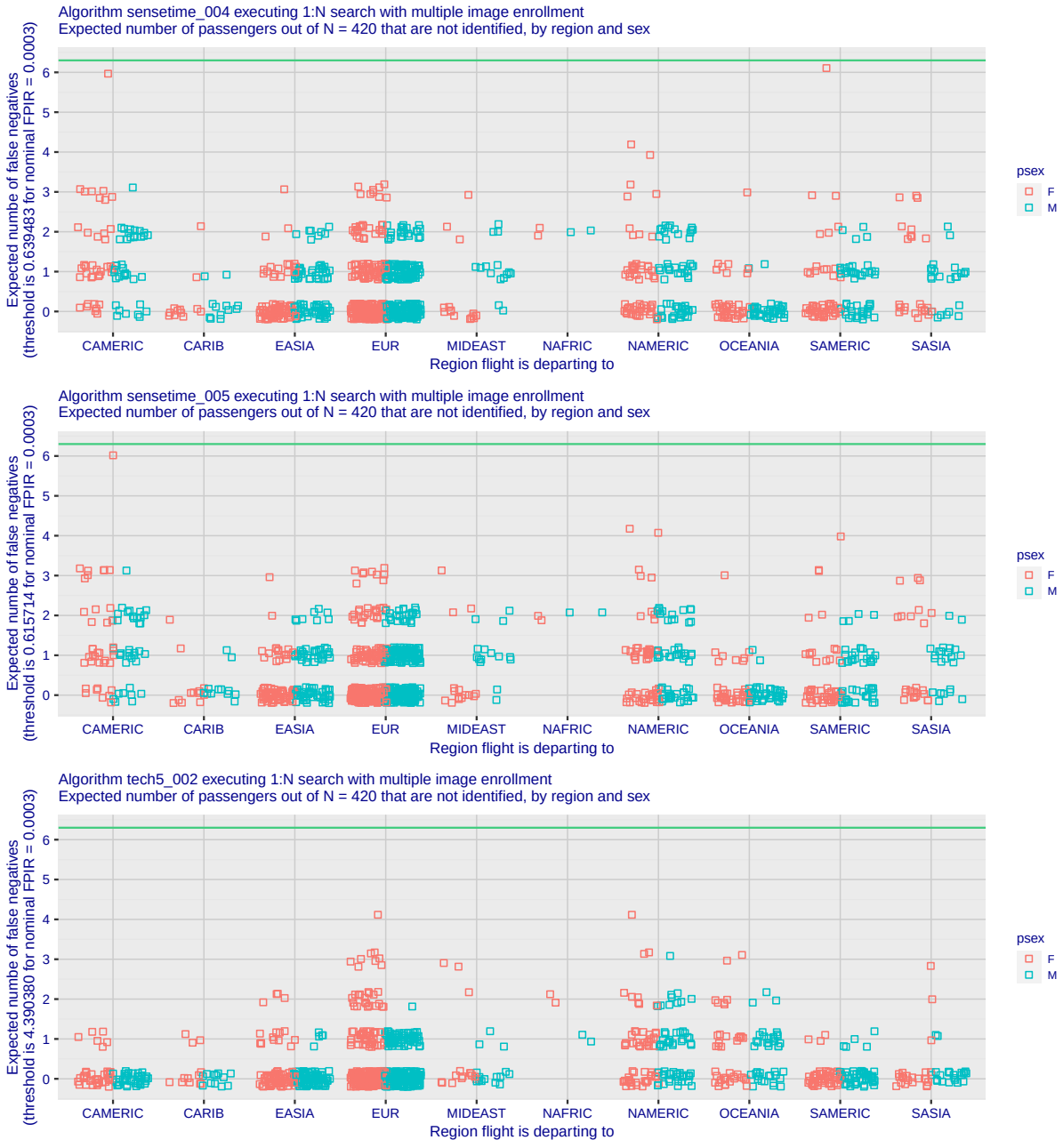


Figure 19: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

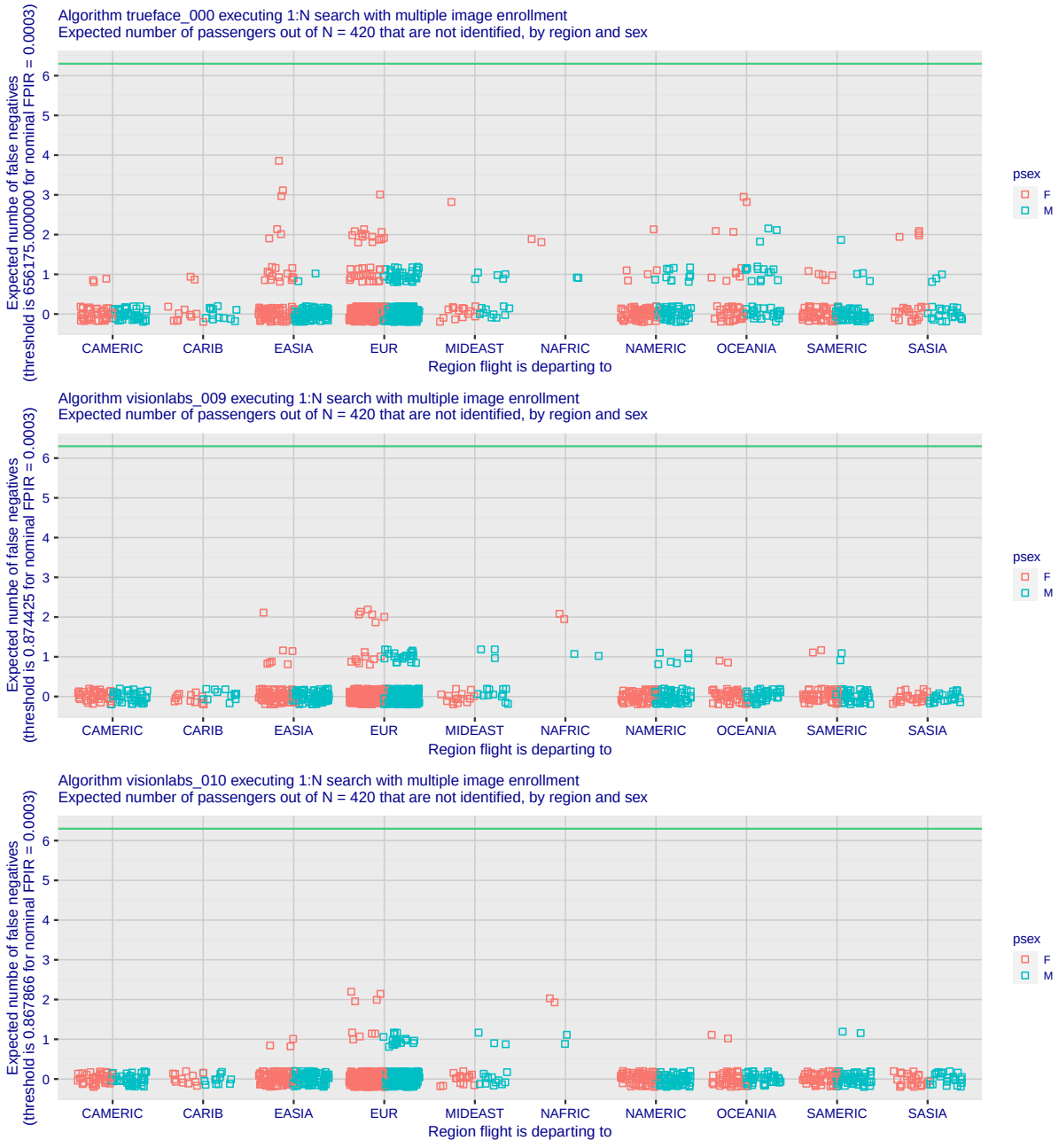


Figure 20: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

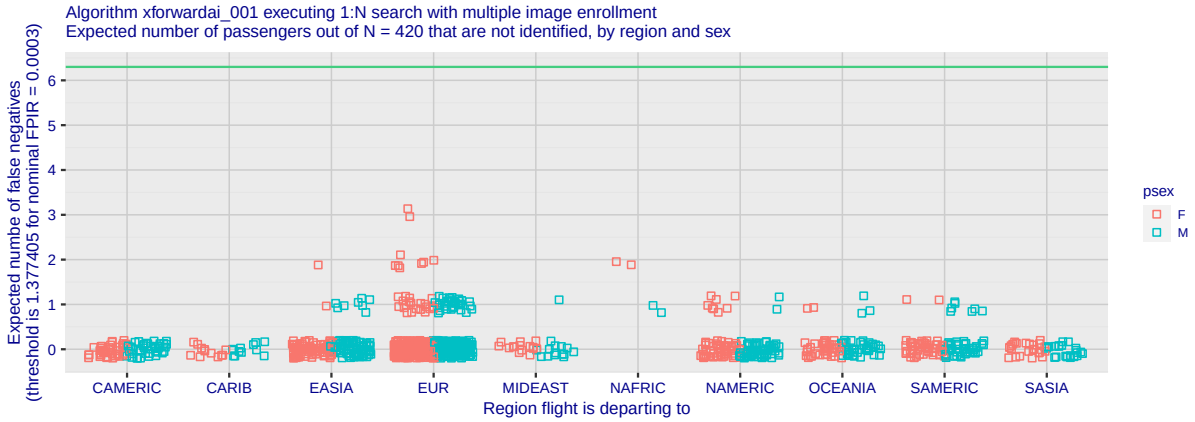


Figure 21: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

Appendix B Figures summarizing false positive identification rate for each algorithm

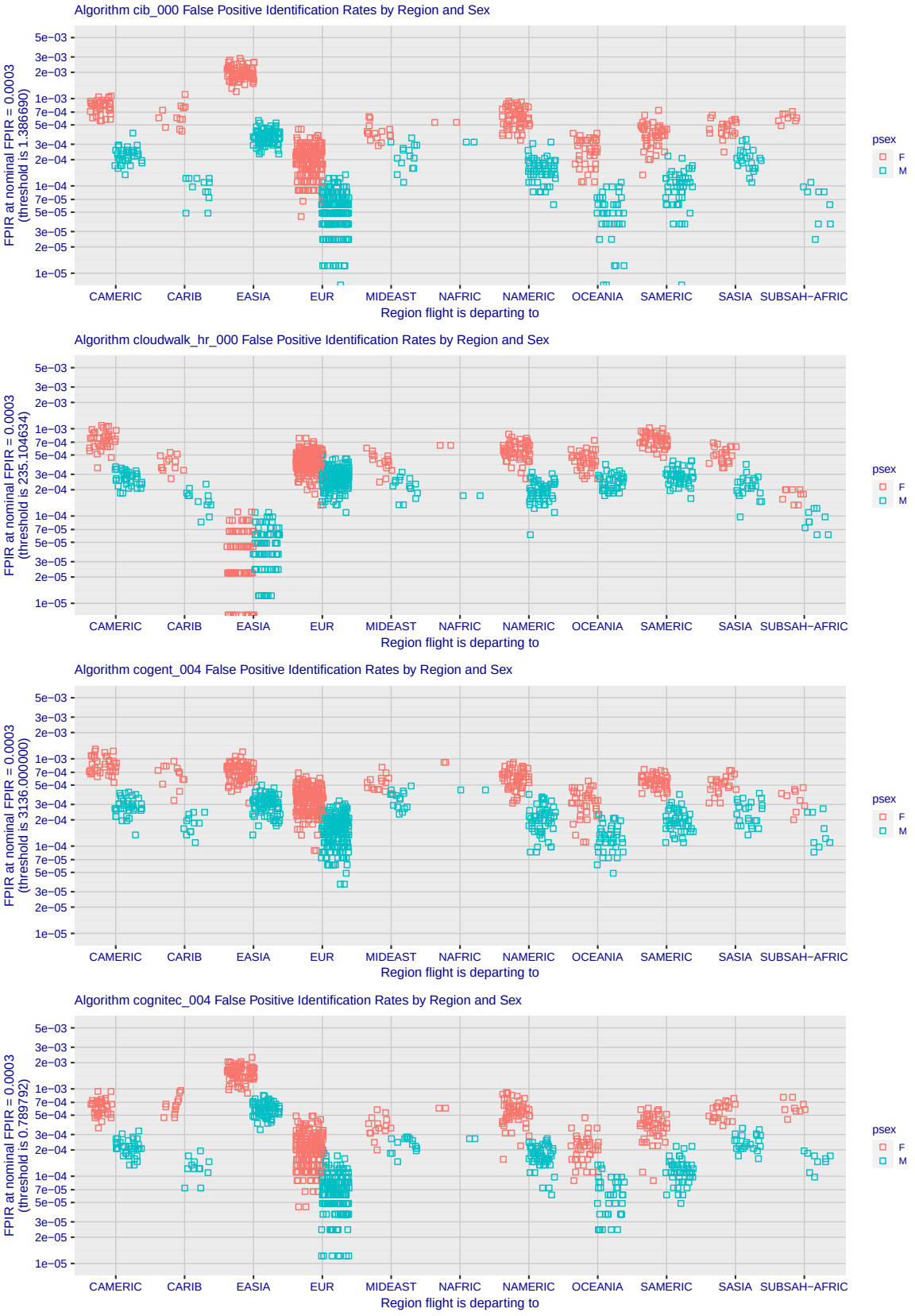


Figure 22: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

Figure 23: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

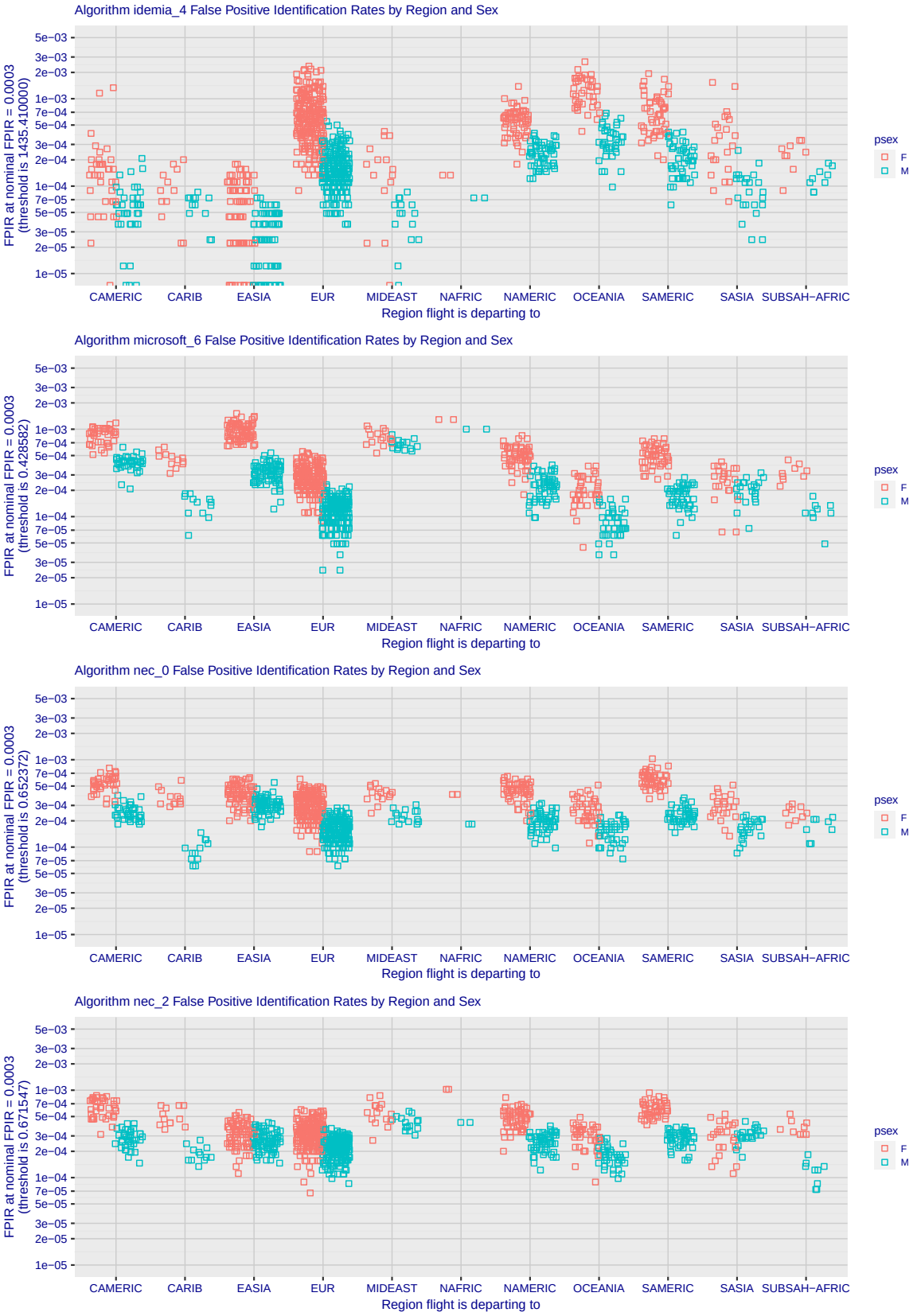


Figure 24: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

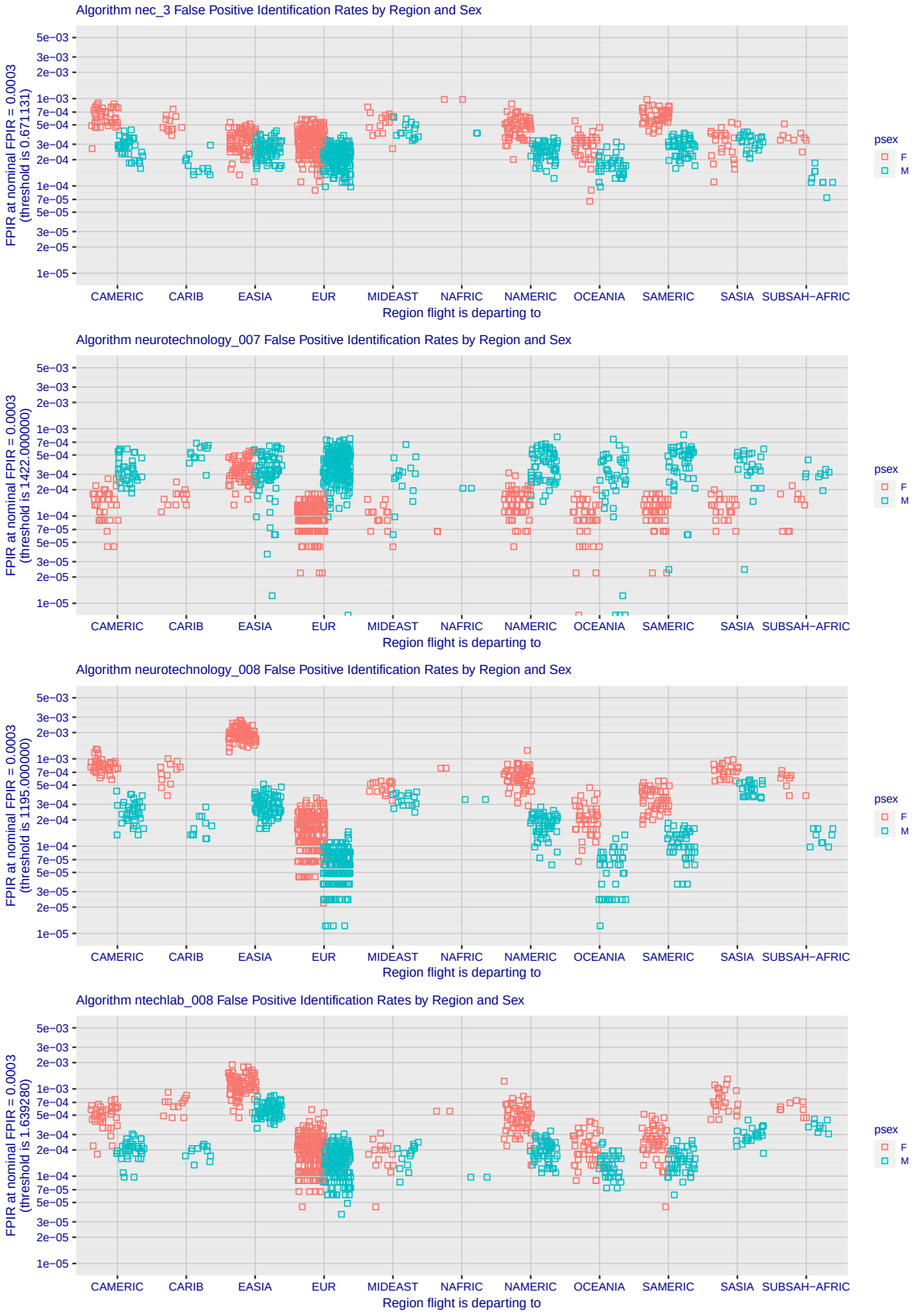


Figure 25: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

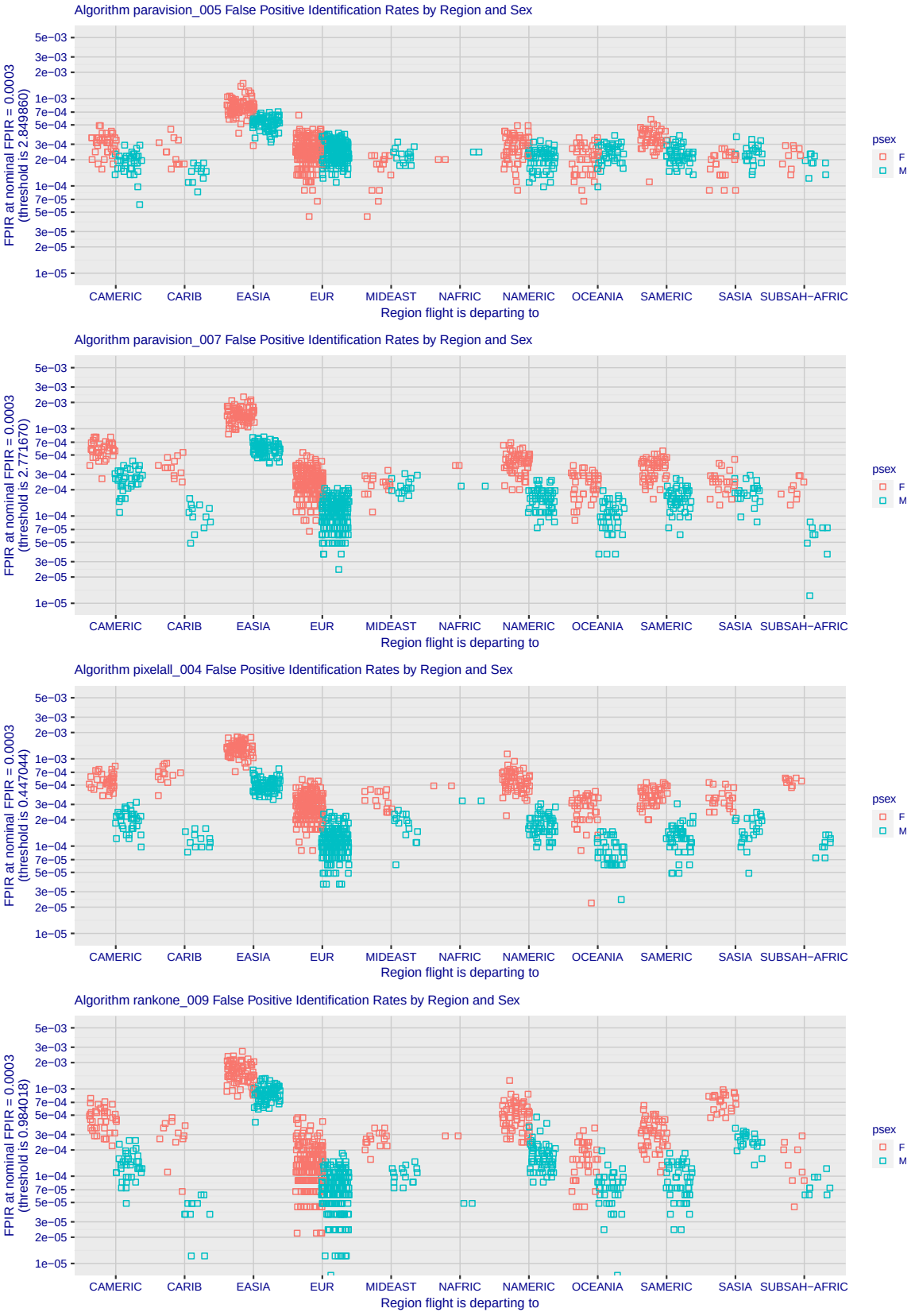


Figure 26: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

Figure 27: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

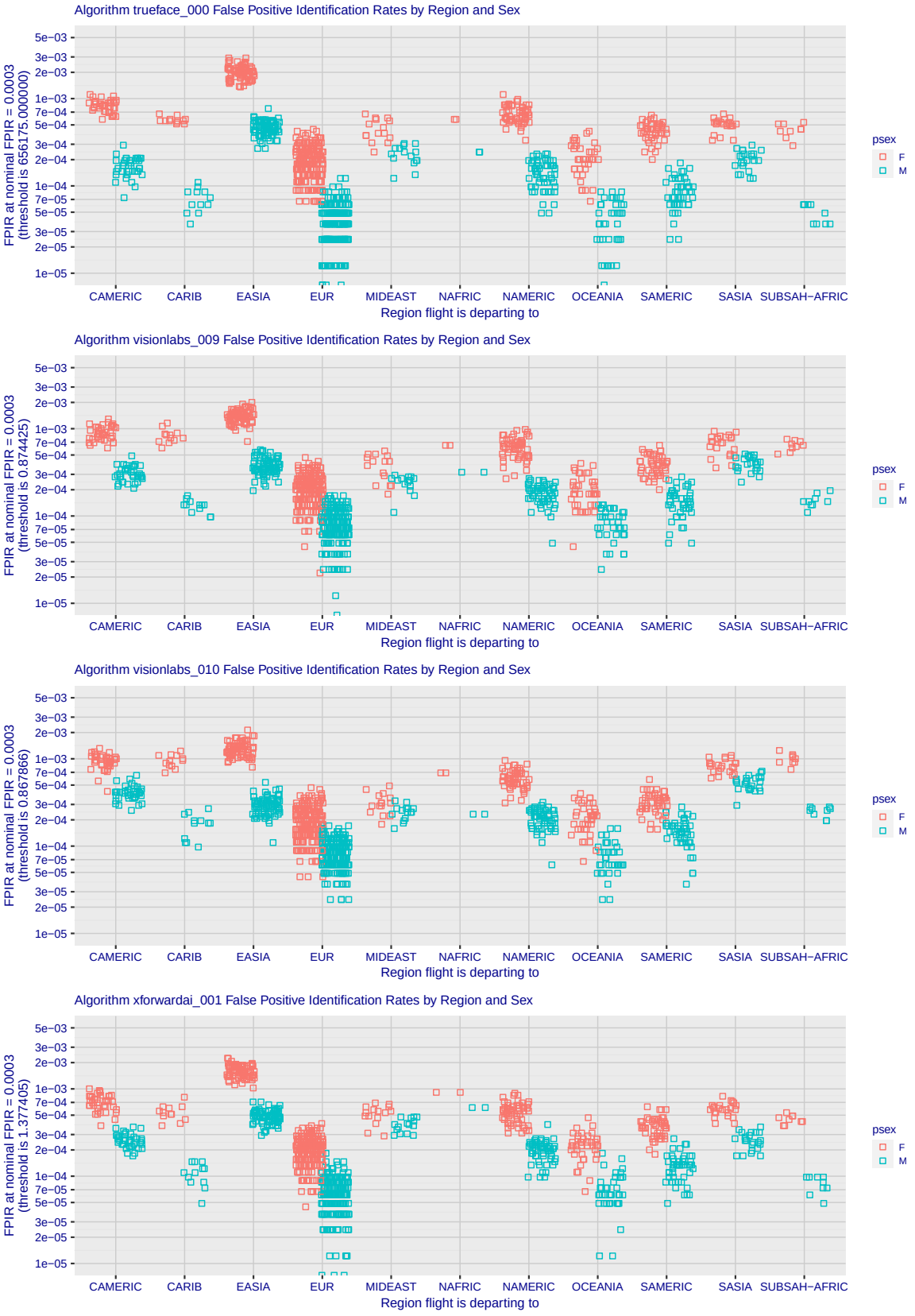


Figure 28: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

References

[1] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 2: Identification. Interagency Report 8271, National Institute of Standards and Technology, Home: https://pages.nist.gov/frvt/html/frvt1N.html, September 2019. https://doi.org/10.6028/NIST.IR.8271.

[2] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Interagency Report 8280, National Institute of Standards and Technology, Home: https://pages.nist.gov/frvt/html/frvt11.html, December 2019. https://doi.org/10.6028/NIST.IR.8280.