Received: 12 April 2022 | Revised: 17 May 2022 | Accepted: 17 June 2022
DOI: 10.1111/1556-4029.15091

ORIGINAL PAPER

Criminalistics

Location distribution of randomly acquired characteristics on a shoe sole

Naomi Kaplan-Damary PhD[1] | **Micha Mandel PhD[2]** | **Yoram Yekutieli PhD[3]** | **Yaron Shor MSc[4]** | **Sarena Wiesner MSc[4]**

[1]Institute of Criminology, Hebrew University of Jerusalem, Jerusalem, Israel
[2]Department of Statistics, Hebrew University of Jerusalem, Jerusalem, Israel
[3]Department of Computer Science, Hadassah Academic College, Jerusalem, Israel
[4]Israel National Police, Division of Identification and Forensic Science (DIFS), Jerusalem, Israel

Correspondence
Naomi Kaplan-Damary, Institute of Criminology, Hebrew University of Jerusalem, Jerusalem 91905, Israel.
Email: naomi.kaplan@mail.huji.ac.il

Abstract

Footwear comparison is used to link between a suspect's shoe and a shoeprint found at a crime scene. Forensic examiners compare the two items, and the conclusion reached is based on class characteristics and randomly acquired characteristics (RACs), such as scratches or holes. An important question concerns the distribution of the location of RACs on shoe soles, which can serve as a benchmark for comparison. This study examines the probability of observing RACs in different areas of a shoe sole using a database of approximately 13,000 RACs observed on 386 outsoles. The analysis is somewhat complicated as the shoes are differentiated by shape and contact surface, and the RACs' locations are subject to measurement errors. A method that takes into account these challenges is presented. All impressions are normalized to a standardized axis to allow for inter-comparison of RACs on outsoles of different sizes and contact areas, and RACs are localized to one of 14 subareas of the shoe sole. Expected frequencies in each region are assumed to be Poisson distributed with rate parameters that depend on the subarea and the contact surface. Three different estimation approaches are studied: a naive crude approach, a shoe-specific random effects model, and an estimate that is based on conditional maximum likelihood. It is shown that the rate is not uniform across the shoe sole and that RACs are approximately twice as likely to appear at certain locations, corresponding to the foot's morphology. The results can guide investigators in determining a shoeprint's evidential value.

KEYWORDS
accidental marks, conditional maximum likelihood, footwear impression, random effects model, randomly acquired characteristics, shoeprints

Highlights

- The distribution of 13,000 randomly acquired characteristics (RACs) on 386 outsoles was evaluated.
- RACs exhibit differences in probability based on their location on the shoe sole.
- This can help to assess the weight of evidence for observed RACs during footwear comparisons.

Presented at the Joint Statistical Meetings (JSM), August 8, 2020, held virtually; the Annual Scientific Meeting of the American Academic of Forensic Sciences, February 17–22, 2020, in Anaheim, CA; and the 104th International Association for Identification Educational Conference, August 11–17, 2019, in Reno, NV.

1 | INTRODUCTION

In recent years, forensic methods have been criticized for their shortcomings in providing courts with objective and quantitative answers to the question of whether a sample from a suspect matches a sample found at the crime scene. Unlike DNA that is used routinely to link suspects to crime scenes because of its scientific objectivity and accessible documentation, the evaluation of pattern evidence such as shoeprints, toolmarks, and even fingerprints has not reached this gold standard. Both the 2009 NRC report, "Strengthening Forensic Science in the United States: A Path Forward," and the 2016 PCAST report to President Obama, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," have called for the strengthening of the scientific basis of forensic procedures [1, 2].

The identification of footwear impressions is based on the comparison of a print found at the crime scene with a test impression made from a suspect's shoe (see [3]), and recent research has focused on scientific methods for this process and ways to analyze its reliability. Kong et al [4] investigated the problem of automatically determining what type (brand/model/size) of shoe left an impression found at a crime scene. Park et al [5] developed an algorithm, which compares two-dimensional footwear outsole images by extracting features that are then combined into a univariate similarity score. Speir et al [6], Richetelli et al [7] and Richetelli et al [8] focused on different aspects of forensic footwear reliability. Here, we study the location distribution of randomly acquired characteristics (RACs) on a shoe sole, which are important for shoe comparison. Figure 1 presents two laboratory prints taken from suspects' shoes, with the location of RACs marked by the examiner. The rarity of this set of RACs is of major interest, especially in establishing a link between a suspect's shoes and the crime scene prints. Thus, the main focus of this study is understanding the spatial distribution of RAC locations, and specifically whether they are distributed uniformly, as assumed by Stone [9], and Wiesner et al [10], or are concentrated in certain areas. This is an essential step in evaluating the degree of rarity of a given set of RACs, that is, the probability that a random shoeprint has a pattern of RACs that is sufficiently similar. Marks at sparsely populated locations would be of much greater value in determining a match than marks at highly populated locations [11–13].

Several studies have considered the probability of RAC locations. In his presentation of a theoretical model that outlines how the consideration of various RAC characteristics could affect the probability of a random match, Stone [9] tentatively proposes a uniform distribution of RACs across the entire shoe sole. Wiesner et al [10] suggest a uniform model for the probability of location using a naive approach described in detail in Section 3. Richetelli et al [11] use a large collection of about 1300 shoes and calculate the probability that RAC's in two independent shoes occur at the same location. A simple adjustment is made to the difference in the contact surfaces in different shoes. The study did not aim to estimate the probability function or create a model that could be extrapolated outside the database and did not utilize information from neighbor locations. Richetelli [14] expands upon this work to explicitly evaluate RAC

FIGURE 1 Locations of RACs marked on lab prints of suspects' shoes (Courtesy of the Israel National Police Division of Identification and Forensic Science – DIFS). [Color figure can be viewed at wileyonlinelibrary.com]

distributions using spatial statistics/spatial modeling. Spencer and Murray [12] use the database of Wiesner [10] and fit a rather complicated spatial hierarchical Bayesian model to the location of RACs, allowing for shoe-specific rate functions. However, they use a local dependence model, which may be less suitable when the location variable is subject to measurement errors.

Similar to Spencer and Murray [12], the RACs' locations here are modeled as a spatial process, and the rate function is estimated in order to calculate the probability of observing RACs in different locations. However, since the location variable is subject to measurement errors and the exact location is not well defined, we replace the continuous rate function with a step-wise constant model over the shoe sole area of interest. This model is fitted to a database of RACs collected by the Israeli Police Division of Identification and Forensic Science, which includes about 13,000 RACs from 386 laboratory shoeprints [10].

Estimation of location distribution arises naturally in the framework of spatial statistics [15], but there are several complications in the current analysis of shoe impressions. First, spatial statistics typically analyzes a single large area having many events or points, while in our case, there are many independent shoe impressions, each with only a small number of events (RACs); about 34 per shoe on average. A second complication is the variability of the shoes: They differ in their types and sizes. Moreover, different shoe soles have different contact surfaces, that is the part of the sole that actually touches the floor or ground and is therefore represented in the shoeprint (see Figure 1, which presents an example of laboratory prints: the area in orange is the contact surface). This fact limits the area in which RACs appear, thus affecting the probability of observing them. On top of these difficulties, some shoe soles are scarred by many RACs, while others have relatively few, apparently due to the level of wear and tear. This article models and estimates the location distribution while taking into account the challenges noted above.

The data used in this article are described in Section 2, and the model is presented in Section 3 along with three estimators of the location distribution; a naive estimator, an estimator based on a random effects model, and an estimator based on conditional maximum likelihood. Section 4 presents an application of the method to the data described in Section 2 and simulation results of a comparison of the three estimators in different settings. Section 5 concludes the paper with a discussion.

2 | DATA

The Israeli Police Division of Identification and Forensic Science (DIFS) has amassed one of the most comprehensive RAC databases, the Jerusalem Shoeprint Accidental Database (JESA), including some 386 laboratory prints having in total more than 13,000 RACs [10]. An important initial preprocessing step was to normalize all shoe impressions to a standardized X-Y axis with identical length and orientation while maintaining the X-Y aspect ratio. This was done by first marking a shoe-aligned coordinate system on each print and then standardizing the shoe sole according to this system: For each laboratory print, the top and bottom of the shoeprint were marked to indicate the direction of the major axis and to determine the length of the shoe sole. As the marking of a shoe-aligned coordinate system is challenging in shoes with a partial contact surface (for example high heels), this was done manually despite the fact that such shoes are less common in the data set which contains mostly sport and casual shoes. The axes' origin was set at the middle point between the two marked extremities. The minor axis was defined as the line perpendicular to the major axis that passes through the origin of the axes. The standardization was done by transforming all measurements from image coordinates to the shoe-aligned coordinate system as follows (for more details, see [10]):

Translation of the marked origin of axes to (0,0).

1. Rotation by the direction of the shoe-aligned coordinate system.
2. Scaling by the length of the shoeprint such that each shoe sole length would be 300 pixels in the down-scaled images.
3. Multiplying the x-value of the points (x being the horizontal axis) by −1 or 1, to mirror whether needed such that all shoeprints would be turned into left shoes.

As the prints are considered here as a two-dimensional representation of the three-dimensional shoe sole, RACs are assumed to have a two-dimensional shape. This study focuses on the location, measured as a point $(x, y) \in \mathbb{R}^2$, which is the center of gravity calculated as the mean of all pixels included in the RAC.

The number of RACs per shoe varies between 1 and 190 with an average of 34 (see Figure S1 in Supplemental Material 1) except for one shoe that has an unusually high number of RACs (309). The RACs were marked by different examiners who were supervised by forensic experts. RACs can be observed only on the contact surface, a feature which varies from one shoe to the other. This should be taken into account in the analysis as described in Section 3. The number of pixels with contact surface per shoe varies between 3631 and 19,199 (out of the total ~122,000 pixels in each down-sampled image, see Figure S2 in Supplemental Material 1). It is also shown that the pad of the shoe sole and four circles at the heel (typical of several shoe models that were common in the data set) more frequently contain a contact surface (see Figure S3 in Supplemental Material 1 for the cumulative contact surface of all shoes). There is a weak correlation, if any, between the number of pixels with contact surface per shoe and the number of RACs per shoe; see Figure S4 in Supplemental Material 1. This suggests that a larger contact surface does not necessarily predict a greater number of RACs.

3 | METHODS

3.1 | Model

As briefly mentioned in the Introduction, using the exact location of RACs is problematic for at least two reasons. First, a RAC is not a point in two dimensions but a set of points, which is marked by trained experts. As the marking process is somewhat subjective and susceptible to error, there is a danger of inaccuracy in the calculation of the RACs' centers. Second, different shoe soles have different shapes, and it is not clear how to adequately normalize them to a standard shoe. Here, the shoe soles were normalized according to the Y-axis, which is the standard shoe size measurement, but the X-axes of different shoes may vary as the original aspect ratio of the shoe sole is preserved. This means that a RAC having a certain X-coordinate can appear near the middle of one shoe sole and near the edge of another. Thus, RAC "locations" are more regional than local, and in order to overcome this, pixels are grouped into larger subsets.

We therefore divide the shoe sole into 14 subareas. On the one hand, these areas should be large enough in order to minimize the errors resulting from the normalization problem, especially on the X-axis. On the other hand, they should reflect the differences in the probability of finding RACs in different parts of the shoe sole, which are mostly attributed to walking patterns. Taking these considerations into account, it was decided that the Y-axis of the shoe sole be divided into 5 layers (separated by $y=-0.35$, $y=-0.15$, $y=0$, $y=0.15$), the X-axis be divided into 2 layers (separated by $x=0$), and the upper part of the shoe sole that comes in contact with the pad of the foot be divided into an outer and inner part as they are expected to behave differently. The final 14 subareas are presented in Figure 2. Subsets denoted by 4 and 11 are separated by the function $y = 12(x+0.02)^2 - 0.07$, subsets 5 and 12 are separated by the function $y = 12(x+0.01)^2 + 0.4$, subsets 6 and 13 are separated by the function $y = 20(x+0.01)^2 + 0.4$, and subsets 7 and 14 are separated by the function $y = 16(x+0.02)^2 - 0.07$. These functions are chosen to fit different types of shoes with different contact surfaces and different contours.

Let i indicate the shoe ($i = 1, \ldots, 386$) and j indicate the subarea ($j = 1, \ldots, 14$). We denote by N_{ij} the number of RACs in shoe i in subarea j and assume that N_{ij} has a Poisson distribution with parameter λ_{ij}. This assumption is quite common in spatial statistics when summarizing count data. In order to reflect the differences in the expected number of RACs between shoes and differences between subareas within a shoe sole, we use the following model:

$$N_{ij} \sim Poisson\left(\lambda_j S_{ij} a_i\right). \quad (1)$$

Here λ_j is the parameter of interest reflecting the relative probability of observing a RAC in subarea j. The shoe-specific parameter a_i enables shoes to have different numbers of RACs reflecting the age and use of the shoe. Finally, S_{ij} is a measured variable denoting the volume of the contact surface in subarea j in shoe i (number of pixels with contact surface).

3.2 | Estimation

There are 14 unknown parameters of interest for the different subareas and 386 nuisance parameters for the shoes. In order to focus on the parameters of interest, we assume a random effects model in which the a_is are independent random variables following a certain distribution with expectation $E(a_i) = 1$. The assumption that $E(a_i) = 1$ is not restrictive, as the Poisson model in Equation (1) is multiplicative and such a constraint is needed so the rate parameters λ_j will be well defined. We consider three estimation approaches and compare them in a later section using a simulation and our shoe database.

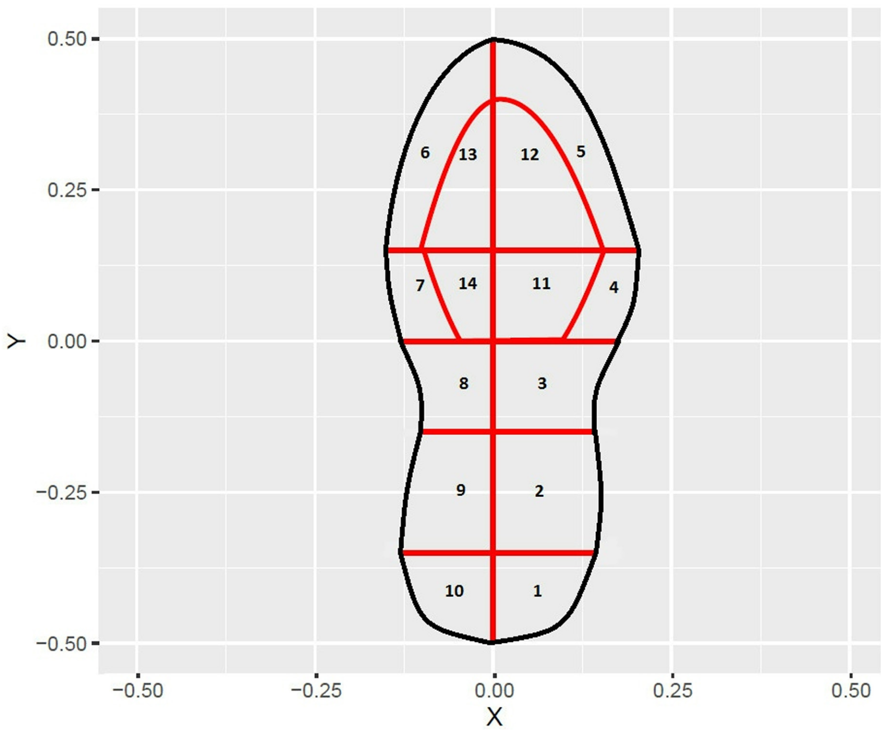

FIGURE 2 Subsets of the shoe obtained according to expert knowledge. [Color figure can be viewed at wileyonlinelibrary.com]

The first naive or moment estimator was previously suggested by Wiesner et al [10]. It is based on the simple observation that, by the model given in Equation (1) and the assumption that $E(a_i) = 1$, if an area has a contact surface (i.e., $S_{ij} > 0$), then $E(N_{ij}|S_{ij}) = \lambda_j$. That is, the expected number of RACs observed in subarea j, standardized by the total contact surface in that area, is exactly λ_j. This gives rise to the following unbiased estimator:

$$\hat{\lambda}_j = \frac{1}{m_j} \sum_i \frac{n_{ij}}{S_{ij}}, \tag{2}$$

where the sum is over all shoes having some contact surface in subarea j, and m_j is the number of such shoes in the database.

The second approach applies a random effects model [16]. Specifically, the model assumes that the shoe effects, a_is are independent and identically distributed following a $Gamma(\theta,\theta)$ distribution. The random effects estimators are obtained in a standard way by maximizing the likelihood with respect to $\lambda_1, \ldots, \lambda_{14}$ and the parameter θ. The *hglm* function under the hglm package [17] in R is used for estimation.

Instead of modeling the distribution of the shoe-specific parameter a_i, the third approach eliminates it by conditioning on its sufficient statistic [18, 19], leading to a conditional maximum likelihood (CML) estimator. Let N_i be the total number of RACs on shoe i. The conditional approach exploits the Poisson property that, given the total number of RACs on a shoe sole, the numbers of RACs on the different areas have a Multinomial distribution. Specifically,

$$N_{i1}, \ldots, N_{i14} \mid N_i = n_i \sim Multinomial\left(n_i, \frac{e^{\log(\lambda_j)+\log(S_{ij})}}{\sum_{j'} e^{\log(\lambda_{j'})+\log(S_{ij'})}}\right)$$

The CML approach reduces to solving the following simple set of equations for $\lambda_1, \ldots, \lambda_{14}$.

$$\sum_{i=1}^{m} \left[\frac{n_{ik}}{\lambda_k} - \frac{n_i}{\sum_{j'} S_{ij'} \lambda_{j'}} s_{ik} \right] = 0 \tag{3}$$

which can be readily carried out numerically. Here, $m = 386$ is the number of shoes in the database.

Under the conditional approach, $\lambda_1, \ldots, \lambda_{14}$ cannot be fully estimated, but only the ratio λ_k / λ_1 which is the relative probability of observing a RAC in subarea k to that of observing a RAC in subarea 1. This replaces the assumption $E(a_i) = 1$, which is not effective under the conditional approach. This should be taken into account when comparing properties of the different methods.

3.3 | The case of a single shoe model

When all shoe soles are similar, the three approaches presented above yield exactly the same estimates. Moreover, this is true even if shoes are of different types, but have the same amount of contact surface in the different subareas. The three methods differ only when not all shoe soles in the database have the same contact surface, and we would expect to see small differences between the approaches in databases containing similar shoe soles. We summarize these results in the following proposition; the proof is deferred to Supplemental Material 2.

Proposition 1 *If* $S_{ij} = S_j > 0$ for all shoes, then

$$\frac{\hat{\lambda}_j}{\hat{\lambda}_1} = \frac{n_{\cdot j}}{n_{\cdot 1}} \cdot \frac{S_1}{S_j}$$

for all subareas j in all three estimators.

As we will see next, when the contact surface varies among the shoe soles, the three approaches can give somewhat different estimates. The simulation study in Section 4.2 compares the three approaches under different settings.

4 | RESULTS

4.1 | Data analysis

The results of the three estimators (naive, random, and CML) applied to the shoe database are presented in Figure 3. Figure 4 presents 95% confidence intervals for the rates in the different subareas of the shoe sole. The interval of the naive estimator was calculated using the normal approximation with variance estimated as described in Section 3 of the Supplemental Material. For the random effects estimator, the *hglm* function under the hglm package [17] was used to calculate the variance. The variance of the CML estimator is based on the observed information matrix. All three estimates agree on the areas with high and low probability.

It is interesting to compare these results with the findings of Gordon [20] who identified the outer heel and area below the first metatarsal head as the most common places for shoe wear to occur as these are the areas where most of the pressure is placed during standing and walking. Hence, assuming that wear, pressure, and RAC probability are correlated, it would be expected to observe a high RAC probability in these areas. Indeed, this is the case in the so-called "stepping circle," subareas 11 and 14. Surprisingly, the outer heel (subarea 1) has the lowest RAC probability. This observation was made earlier by Davis and DeHaan [21] who explained that erosion at the rear heel due to continued wear erased RACs at that position. Thus, extensive wear could theoretically cause opposite phenomena, the creation of multiple RACs, or their eradication.

In order to further understand this phenomenon, the size distribution of RACs in the various shoe sole areas derived from the JESA database used for this research was examined. Table S1 in Supplemental Material 1 presents RAC size quantiles and means by subareas of the shoe sole (1–14). The mean size of the RACs is largest in subarea 1, apparently because of the great wear and tear of this area during walking. The result shows major tearing in this area

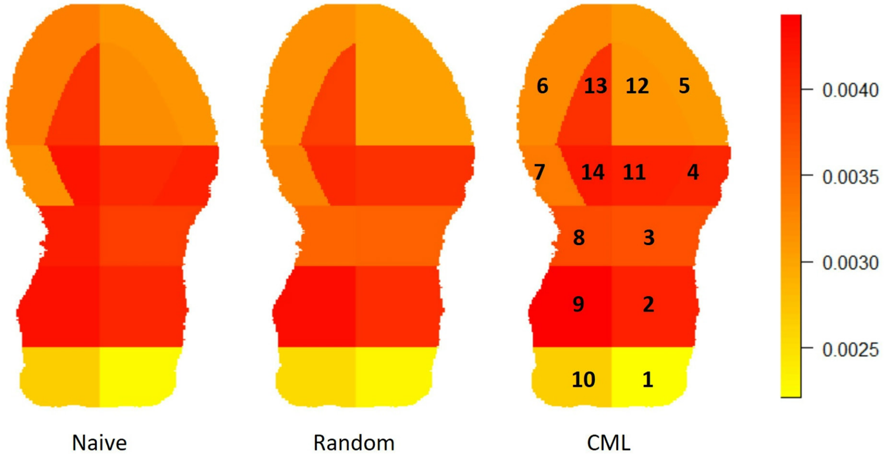

FIGURE 3 The estimated rate function - a comparison of three approaches. [Color figure can be viewed at wileyonlinelibrary.com]

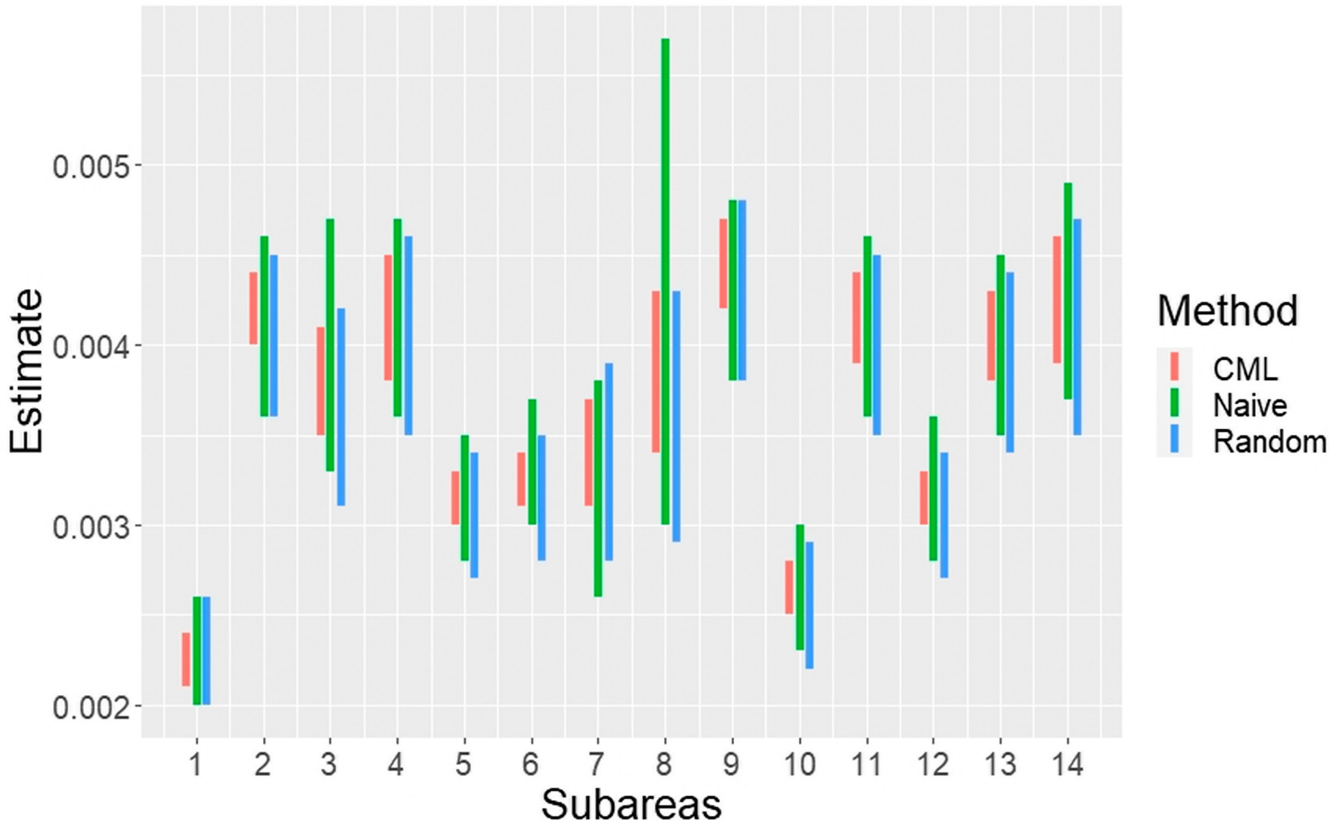

FIGURE 4 Confidence intervals for the estimators based on the piece-wise constant model; values are multiplied by 1000 to simplify presentation. [Color figure can be viewed at wileyonlinelibrary.com]

but numerically a smaller number of RACs than in other areas of the shoe sole. The relatively small probability of RACs at the front rim (subareas 5 and 6) and the outer toe area (subarea 12) is explained by the fact that less pressure is put on these areas during walking, resulting in less scarring of the shoe sole. The number of RACs found in the inner step (subareas 3 and 8) is small. However, as the estimation of location probability takes into account both the number of RACs and the contact surface, the adjusted probability is quite large due to the relatively limited contact surface in the inner step (see Figure S3 in the Supplemental Material 1).

The confidence intervals of the three approaches are relatively close. The confidence interval of the CML approach is narrower as a result of conditioning on the n_js and treating the scaling factor as a constant. The estimates agree on the areas with relatively wide and narrow intervals. The widest interval is of subarea 8 of the shoe sole, which is characterized by a low amount of contact surface (see Figure S3 in Supplemental Material 1). In addition, using the random effects model, the hypothesis that the λ_j parameters are equal for all j, meaning that the rate is uniform over the whole shoe sole, is rejected with a $p - value \approx 0$. Thus, there is a clear deviation from a

JOURNAL OF FORENSIC SCIENCES

uniform model, assumed by Stone [9] and Wiesner et al [10], and the maximum estimated rate value is about twice that of the minimum value. In spite of the fact that there is no area in which observing a RAC can increase dramatically the evidential value of the shoe, observing multiple RACs may do so.

4.2 | Simulation

The random effects and CML approaches give almost identical results when applied to the real database, with minor differences from the naive estimates (see Figure 4). In this section, we further compare the methods and evaluate their properties by simulation. The simulation study compares the three estimators based on parameters from the shoe data, assuming that the number of RACs follows the model in Equation (1). The first simulation uses the naive estimates reported in the previous subsection for the subarea parameters λ_j, and the observed contact surfaces, S_{ij} in the 14 subareas. The number of shoes in the simulation is 386, as in the original data set. In each replication of the simulation, the shoe-specific parameters a_i are simulated from a $Gamma(\theta,\theta)$ distribution, where $\theta = 0.908$ is the estimate of $1/Var(a)$ based on the data (see Supplemental Material 3). The results are based on 500 replications. Figure 5 presents the relative bias, which is the empirical bias of the estimator divided by the real parameters' value and the ratio between the estimators' MSE and the theoretical variance of the naive estimator (see Supplemental Material 3). The naive estimator is unbiased, and the simulation suggests that the biases of the two other estimators are negligible in this setting. The random effects estimator has the lowest relative bias and MSE across all parameters. In addition, the naive estimators' MSE ratio is around 1, and since it is unbiased, this indicates that its theoretical variance is close to the empirical variance.

Additional simulations which investigate the effect of λ, the effect of the number of subareas, the effect of the sample size, the number of shoes, and the effect of a_i are conducted. The specific settings and the results appear in an unpublished thesis [22]. In summary, the random effects estimator is found to be the best among the estimators in most settings, with good performance in all. The CML estimator is very close to it.

5 | DISCUSSION

The findings of this study can help in assessing the potential evidential value of shoeprints. It is clear that RACs in certain locations are rarer than others, and taking this into account may contribute to an assessment of the rarity of a given shoe sole. In the past, whether consciously or not, examiners treated all RAC locations as equally probable. As the rate function provides an indication of the RACs' degree of rarity, it is important in assessing the RAC's evidential value and can serve as a supporting tool for the examiner in the evaluation of shoeprints.

Other characteristics of the RAC, such as size and shape, can be studied to further improve the classification of RACs as common or rare, but these are prone to severe measurement errors that must be modeled. Additional challenges confront the researcher when analyzing crime scene prints, which are complicated by noise of various forms. Further research on actual crime scene RACs and their comparison with laboratory prints is therefore of vital importance.

This study is limited by the nature of the data on which it is based. The data set was collected by the Israel Police over a period of ten years from real suspects and thus represents an authentic collection of relevant shoes. Apparently, certain types of sport shoes were more common, and this could affect our results given the frequency

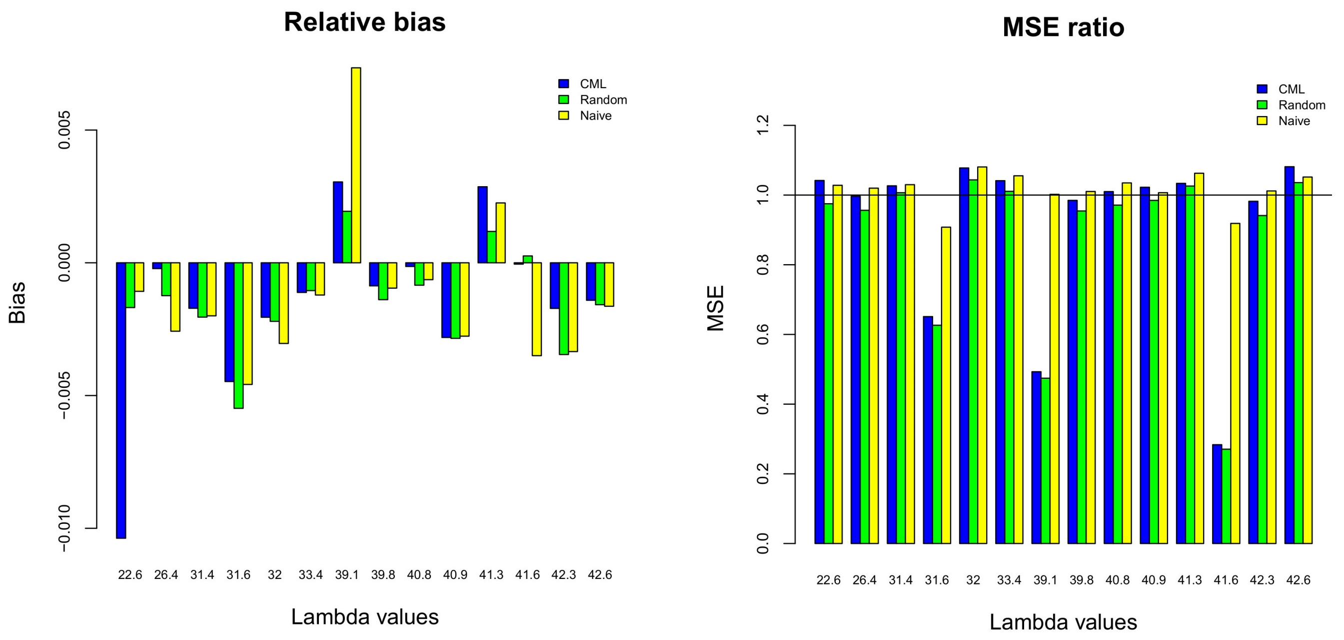

FIGURE 5 The relative bias and MSE ratio. [Color figure can be viewed at wileyonlinelibrary.com]

of outsoles with certain designs. Casework containing other styles could lead to different findings.

There are limitations in using the 14-subarea partition chosen for this analysis, which was guided by professional experience rather than empirical data. Different partitions would naturally represent the findings differently: larger subsets would reduce the resolution and possibly mix areas with high and low probability of finding RACs, whereas smaller areas are prone to the problem of different shoe sole shapes. Different shapes and locations of subareas could also have a significant impact. Our findings are consistent with the results of an alternative high-resolution partition produced by a pixel-based model (see [23]), which employs a logistic regression with natural cubic splines for the location variable, and yields a smooth intensity function. However, this latter pixel-based model itself poses some challenges. As noted in Section 3.1, the definition of location is problematic. Richetelli et al [24] and Yekutieli et al [25] estimated that the magnitude of the location error is on the order of 5 mm when repeating the marking process on a single shoe. Yekutieli et al [25] further notes that this error was found to result largely from the error in assigning the shoe-aligned coordinate system, and to a lesser extent from the error in estimating the centroid of the RACs. There is an additional source of error which was not noted in their study. Normalization of shoe soles along the Y-axis alone (and not the X-axis in order to preserve the original aspect ratio) fails to eliminate differences between wide and narrow shoe soles as shown in Figure S5 in Supplemental Material 1. This is expected to be substantial in defining the X-axis location. For this reason, we take a cautious stance regarding the accuracy of location information and assume that location error could be greater. Thus, "locations" of RAC's are more regional than local, and pixels were therefore grouped into larger subsets. The optimal way to divide the shoe sole and to determine the subsets remains and should be addressed in future research.

The question of the shoe sole partition also arises with regard to the interpretation of estimated rates in different subareas. There seems to be evidence that certain areas have similar estimated rates based on their confidence intervals (subareas 5, 6, and 12, subareas 2 and 4, and subareas 11, 13, and 14). However, inference based on these confidence intervals is post hoc and thus may be biased.

The estimation of the rate function has been made under the assumption of independence among RACs. However, as shown by Kaplan-Damary et al [26], this assumption is unjustified. This does not invalidate the findings, as using an independence working assumption results in consistent estimators, but the variance estimator may be somewhat biased. Using larger areas as building blocks for the model may solve part of the local dependence problem, but further study is needed to understand the importance of the assumption.

The CML and random effects approaches produced very similar results, which were relatively close to the naive estimates. The simulations suggest that, among the three approaches, the random effects estimator performs best and thus may be preferred. Using this model, the hypothesis of a uniform probability over the shoe sole is rejected, and the maximum estimated value is approximately twice that of the minimum value. The estimated probability is the smallest at the toes and heel of the foot. The deviation from uniformity is likely to be a result of the morphology of the foot and the areas of the foot that cause pressure on the shoe. This fits the shape of the estimates presented here.

ORCID

Naomi Kaplan-Damary https://orcid.org/0000-0001-8218-7654
Micha Mandel https://orcid.org/0000-0002-4270-9777
Yoram Yekutieli https://orcid.org/0000-0002-4398-4906

REFERENCES

1. National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press; 2009. https://doi.org/10.17226/12589
2. Executive Office of the President President's Council of Advisors on science and technology. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: PCAST; 2016.
3. Bodziak WJ. Footwear impression evidence: detection, recovery and examination. 2nd ed. Boca Raton, FL: CRC Press; 2000. p. 357–72.
4. Kong B, Supancic J, Ramanan D, Fowlkes C. Cross-domain forensic shoeprint matching. Proceedings of the 2017 British machine vision conference (BMVC 2017). London, UK, Ames, IA: Iowa State University; 2017, 2020. p. 1–5. [cited 2022 June 6]. Available from: https://dr.lib.iastate.edu/server/api/core/bitstreams/bee8564e-4ad1-4155-ace8-9da1f48a6c17/content
5. Park S, Carriquiry A. An algorithm to compare two-dimensional footwear outsole images using maximum cliques and speeded-up robust feature. Stat Anal Data Min: ASA Data Sci J. 2020;13(2):188–99. https://doi.org/10.1002/sam.11449
6. Speir JA, Richetelli N, Hammer L. Forensic footwear reliability: part I – participant demographics and examiner agreement. J Forensic Sci. 2020;65(6):1852–70. https://doi.org/10.1111/1556-4029.14553
7. Richetelli N, Hammer L, Speir JA. Forensic footwear reliability: part II – range of conclusions, accuracy, and consensus. J Forensic Sci. 2020;65(6):1871–82. https://doi.org/10.1111/1556-4029.14551
8. Richetelli N, Hammer L, Speir JA. Forensic footwear reliability: part III – positive predictive value, error rates, and inter-rater reliability. J Forensic Sci. 2020;65(6):1883–93. https://doi.org/10.1111/1556-4029.14552
9. Stone RS. Footwear examinations: mathematical probabilities of theoretical individual characteristics. J Forensic Identif. 2006;56(4):577–99.
10. Wiesner S, Shor Y, Tsach T, Kaplan-Damary N, Yekutieli Y. Dataset of digitized RACs and their rarity score analysis for strengthening shoeprint evidence. J Forensic Sci. 2019;65(3):762–74. https://doi.org/10.1111/1556-4029.14239
11. Richetelli N, Bodziak WJ, Speir JA. Empirically observed and predicted estimates of chance association: estimating the chance association of randomly acquired characteristics in footwear comparisons. Forensic Sci Int. 2019;302:109833. https://doi.org/10.1016/j.forsciint.2019.05.049
12. Spencer NA, Murray JS. A Bayesian hierarchical model for evaluating forensic footwear evidence. Ann Appl Stat. 2020;14(3):1449–70. https://doi.org/10.1214/20-AOAS1334
13. Speir JA, Richetelli N, Fagert M, Hite MN, Bodziak WJ. Quantifying randomly acquired characteristics on outsoles in terms of shape and position. Forensic Sci Int. 2016;266:399–411. https://doi.org/10.1016/j.forsciint.2016.06.012

14. Richetelli N. Statistical evaluation of randomly acquired characteristics on outsoles with implications regarding chance co-occurrence and spatial randomness [dissertation]. Mogantown, WV: West Virginia University; 2020.
15. Cressie N. Statistics for spatial data. New York, NY: Wiley; 1993. p. 7–10. https://doi.org/10.1002/9781119115151
16. Myers RH, Montgomery DC, Vining GG, Robinson TJ. Generalized linear models: with applications in engineering and the sciences. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2010. p. 319.
17. Rönnegård L, Shen X, Alam M. Hglm: a package for fitting hierarchical generalized linear models. The R Journal. 2010;2(2):20–8.
18. Bishop YM, Fienberg SE, Holland PW. Estimating the size of a closed population. Discrete multivariate analysis: theory and practice. New York, NY: Springer; 2007. p. 229–56. https://doi.org/10.1007/978-0-387-72806-3_6
19. Agresti A. Categorical data analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2013. p. 265–6.
20. Gordon TL. Corrective treatment of the foot. Chiropodist. 1940;27(244):159–66.
21. Davis RJ, DeHaan JD. A survey of men's footwear. Sci Justice. 1977;17(4):271–85.
22. Kaplan-Damary N. Statistical methods for evaluating forensic evidence [dissertation]. Jerusalem, Western Asia: Hebrew University of Jerusalem; 2019.
23. Kaplan-Damary N, Mandel M, Yekutieli Y, Wiesner S, Shor Y. Spatial modeling of randomly acquired characteristics on outsoles with application to forensic shoeprint analysis. arXiv:1912.08272v1 [Preprint]. 2019. p. 45. [cited 2022 June 29]. Available from: https://arxiv.org/abs/1912.08272v2
24. Richetelli N, Nobel M, Bodziak WJ, Speir JA. Quantitative assessment of similarity between randomly acquired characteristics on high quality exemplars and crime scene impressions via analysis of feature size and shape. Forensic Sci Int. 2017;270:211–22. https://doi.org/10.1016/j.forsciint.2016.10.008
25. Yekutieli Y, Shor Y, Wiesner S, Tsach T. Expert assisting computerized system for evaluating the degree of certainty in 2D shoeprints. Final report. National Institute of Justice: Washington, DC; 2016 Report No.: TP-3211.
26. Kaplan-Damary N, Mandel M, Wiesner S, Yekutieli Y, Shor Y, Spiegelman C. Dependence among randomly acquired characteristics on shoeprints and their features. Forensic Sci Int. 2018;283:173–9. https://doi.org/10.1016/j.forsciint.2017.11.038

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kaplan-Damary N, Mandel M, Yekutieli Y, Shor Y, Wiesner S. Location distribution of randomly acquired characteristics on a shoe sole. J Forensic Sci. 2022;67:1801–1809. https://doi.org/10.1111/1556-4029.15091