Testing Machine Learning Fairness Applied to RaspBary

Ryan Hosler
Indiana University-Purdue University Indianapolis,
420 University Blvd, Indianapolis, IN 46202
rjhosler@iu.edu

***Abstract*— RaspBary is a lite weight machine learning python application for urban logistics. Before its deployment in the spring, a fairness assessment ought to be done to assure non-discriminatory behavior. This paper introduces the details of the RaspBary application and current metrics for testing machine learning fairness. When applying those tests to simulated trials, RaspBary shows no statistically significant discriminatory behavior.**

Keywords*— *Fairness, Hawke's process, Wasserstein barycenter, machine learning, disparate impact.

I. Introduction

The first real-world use of RaspBary (Real-time Allocation Service with Point process BARYcenters) will be for optimizing the Marion county EMS department's ambulance response time. Faster response time has potential for improving public health. As discussed in [3], faster response times have been correlated to an increased survival rate for out-of-hospital cardiac arrests. Since this is a sensitive use for the algorithm, some due caution should be applied. The goal of this paper is to ensure that the application does not have indirect discriminatory behavior to protected minority classes. To determine parity, the average ambulance distance will be tested through disparate impact. Before tests can be conducted, background knowledge of RaspBary and disparate impact must be explored.

II. Background

Currently, integrating scalable Hawke's process software with mobile web applications remains a challenge (see [10] for a review of spatio-temporal Hawke's process literature). RaspBary is an application that intends to provide this service through modular api calls. The api-server interface is demonstrated in figure 1 (code is available in a public github repository [7]):

- A GET request will return intensity data when given a projection time window
- A POST requests will return optimal vehicle locations when given current vehicle locations
- A GET request will update the model given a batch of CSV events.

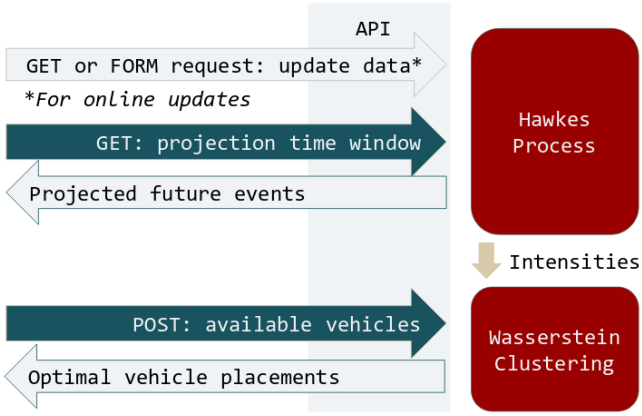

Fig. 1 RaspBary interface diagram from [6].

A. Hawke's Process.

Shown in figure 2, the Hawke's process model utilizes an exponential basis for the triggering function. A non-parametric method is used to estimate hourly trends, day-of-week trends, and monthly trends. A grid of intensities is provided to allow for sorting and organizing when and where an event is likely to occur. Intensity rates per grid cell is shown in figure 3. This implementation allows for online gradient descent training and privacy preservation [8].

$$\lambda_g(t) = \mu_g + \sum_{k=1}^{K} \sum_{\substack{t>t_j \\ g_j=g}} \theta_k \omega_k e^{-\omega_k (t-t_j)}$$

Fig. 2 Hawke's process triggering function

$$\lambda_{Final\_g}(t) = \lambda_g(t) M(t) D(t) H(t)$$

Fig. 3 Final intensity in each grid cell

The focus for fairness testing in this paper regards EMS data. When requested, the RaspBary application will return a gridded heatmap of predicted emergencies for a given time interval. Demonstrated in figure 4, this data allows for an estimated number of total emergencies that will occur.

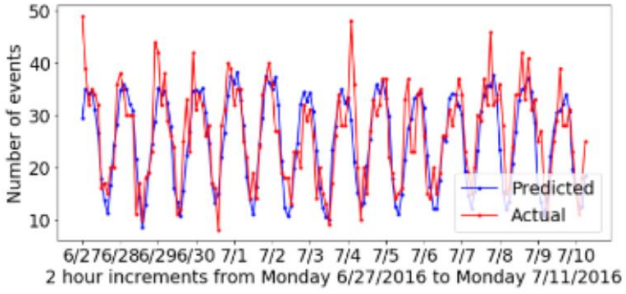

Fig. 4 RaspBary predicting total number of emergencies.

B. Wasserstein Barycenters.

For vehicle allocation, Raspbary returns wasserstein barycetners to point process intensity [6]. These barycenters minimize the function in figure 5 to optimally allocate vehicles. As a result, the wasserstein distance between vehicles over an intensity map is minimized. For algorithmic efficiency, these are calculated by the fast approach given by Cuturi [2].

$$f(a, X) \stackrel{\text{def}}{=} \sum_{i=1}^{N} p(a, b_i, M_{XY_i})$$

Fig. 5 Final intensity in each grid cell

C. Disparate Impact.

Disparate impact (DI) is a statistical measurement for comparing the proportional probability of a success between different populations. A success is determined by a binary classifier where a 1 is defined as a success and 0 is a failure. As shown in figure 6 [1], *S* corresponds to the demographic type of the population. For example, imagine an algorithm that chooses who to hire from a set of applicants. A Y = 1 indicates a "hired" and a Y = 0 indicates a "not hired." In this scenario, the algorithm should not exclusively hire men over women. In this case, women is assumed to be the unfavored clas S = 0 and men is the favored class S = 1.

$$DI(Y) := \frac{\mathbb{P}(Y = 1 | S = 0)}{\mathbb{P}(Y = 1 | S = 1)}$$

Fig. 6 Disparate Impact statistic

For interpreting DI, a value of 1 indicates perfect parity; hence, the probability of a successful outcome will be the same for the favored and unfavored class. Consequently, the closer DI is to 0, the more unfair the algorithm is to unfavored class. As cited by [5], a DI of 0.8 is indirectly the legal score for determining if an algorithm is unfair. The notation for this score will be $\tau = 0.8$. For the sake of testing RaspBary, the tail-end alternative hypothesis for statistically significant DI will be $H_a < \tau$. The null hypothesis will then be $H_0 = \tau$.

The goal of this paper is to determine if bias, not parity, is statistically significant. That is why the null hypothesis is set to the legal score of τ.

D. Related Work

Determining algorithmic parity for self-exciting Hawke's process algorithms has been a concern in the past. For example, in [2], bias in a predicitive policing algorithm was tested. The main concern: the algorithm could encourage police tactics that target minorities as a method for reducing crime. The data gathered in this paper came from a randomized predicive police trial. Real world data indicated that the predictive policing algorithm did not induce more bias than normal patrolling. To estimate a trail and analysis similar to [2], simulated use of RaspBary will be conducted over real emergency data.

III. Methods

For testing the fairness of the application, its real-world use will be simulated. Since the application intends to reduce ambulance response time, the distance to an emergency is the metric that will be compared. Moreover, DI confidence intervals will be used to provide a more complete statistical analysis.

A. Method for Data Retrieval

Gathering data such that a fairness assessment is possible will follow the following steps:

1. Get a gridded map of emergency intensities for the next two hours.
2. Allocate several ambulances over that grid to optimize response time.
3. Use the next real emergency that occurred over the given time period.
4. Using the geographical coordinate of the real emergency, randomly sample that block's demographics to guess the race of the emergency.
5. Compute the distance from the nearest ambulance to that emergency.
6. Once 2 hours of sequential emergencies have occurred, repeat 1 – 6 until a desired a amount of simulated data is reached

As described by the Marion County EMS department, they intend to allocate ambulances with RaspBary during peak event time. This requirement is why the emergency data estimates total number of emergencies for a given time period (shown in figure 4). The simulation described here demonstrates how a full utilization of RaspBary would work.

Since the DI of no utilization of RaspBary will act as the control for this experiment, a "utilization of RaspBary during peak event time" data set will not be used. Attempting to simulate that usage of RaspBary would have some flaws. First, there is not a defined number of emergencies that would trigger a usage of RaspBary. Second, a lieutenant controls the number of ambulances that get allocated outside of EMS stations. Attempting to simulate these two characteristics would require human heuristic estimation. Since the data of the EMS department using RaspBary does not exist, such a simulation is not feasible.

B. Method for Applying Disparate impact.

DI comparisons requires a binary label. For this reason, a classification value will be added to the data:

- 1: Distance to emergency < median distance to emergency
- 0: otherwise

This classification simply describes a success as a response distance less than the median response distance of the simulated data. Now it is possible to compare the number of "short distances" across demographic variables.

For applying confidence intervals to this classification metric, the methods given by [1], shown in figure 7, will be used.

$$T_n := \frac{\sum_{i=1}^{n} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{S_i=0} \sum_{i=1}^{n} \mathbb{1}_{S_i=1}}{\sum_{i=1}^{n} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{S_i=1} \sum_{i=1}^{n} \mathbb{1}_{S_i=0}}.$$

Then the asymptotic distribution of this quantity is given by

$$\frac{\sqrt{n}}{\sigma}\left(T_n - DI(g,X,S)\right) \xrightarrow{d} N(0,1), \text{ as } n \to \infty,$$

where $\sigma = \sqrt{\nabla\varphi^T(\mathbb{E}Z_1)\,\Sigma_4\,\nabla\varphi(\mathbb{E}Z_1)}$ *and*

$$\nabla\varphi^T(\mathbb{E}Z_1) = \left(\frac{\pi_1}{p_1\pi_0}, -\frac{p_0\pi_1}{p_1^2\pi_0}, -\frac{p_0\pi_1}{p_1\pi_0^2}, \frac{p_0}{p_1\pi_0}\right)$$

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0p_1 & p_1(1-p_1) & & \\ \pi_1p_0 & -\pi_0p_1 & \pi_0\pi_1 & \\ -\pi_1p_0 & \pi_0p_1 & -\pi_0\pi_1 & \pi_0\pi_1 \end{pmatrix},$$

where we have denoted $\pi_s = \mathbb{P}(S_1 = s)$ *and* $p_s = \mathbb{P}(g(X_1) = 1, S_i = s)$, $s = 0, 1, .$

Fig. 7 Statistics Used for computing Confidence Intervals.

Now confidence intervals for DI can simply be done by applyning the equation $(T_n \pm \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}})$ given by [1]. For fairness testing, the 95% confidence level will be used.

IV. Data and Results

The response data for RaspBary is simulated over 5000 real emergency points. For applying fairness, the three largest demographics of white, black, and hispanic are compared. Since there are 25 EMS stations in Marion County, a full RaspBary usage simulation will initially be done with 25 ambulances. Data from this initial trail is demonstrated in fiqure 8. Since RaspBary allows for a variable number of allocated ambulances, its full usage for 15 and 35 ambulances are shown in figure 9 and figure 10 respectively. Lastly, the control data of response distances just from the current EMS stations are shown in figure 11. These graphs use the normal 1.5 IQR range for displaying data

```
Populations:  White: 2750 Black: 1740 Hispanic: 510
Total time passed:  16 days 13:39:49

white:  median:  0.983 mean:  1.214
black:  median:  0.933 mean:  1.052
hispanic:  median:  1.001 mean:  1.123
```

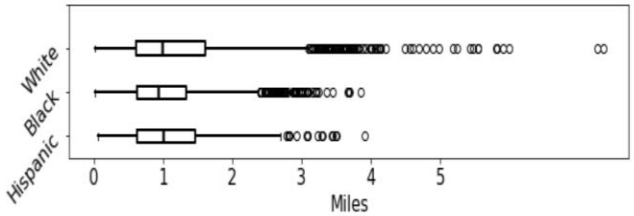

Fig. 8: Response Distance by Demogrpahic for 25 ambulances

```
Populations:  White: 2774 Black: 1705 Hispanic: 521
Total time passed:  16 days 13:39:49

white:  median:  0.874 mean:  1.134
black:  median:  0.82 mean:  0.976
hispanic:  median:  0.856 mean:  1.067
```

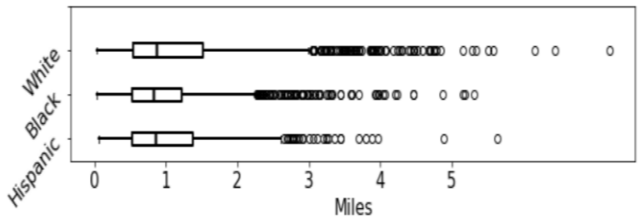

Fig. 9: Response Distance by Demogrpahic for 35 ambulances

```
Populations:  White: 2738 Black: 1771 Hispanic: 491
Total time passed:  16 days 13:39:49

white:  median:  1.301 mean:  1.557
black:  median:  1.199 mean:  1.33
hispanic:  median:  1.219 mean:  1.405
```

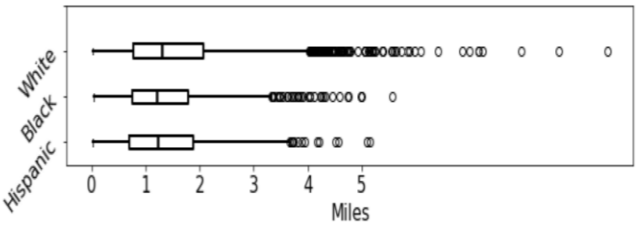

Fig. 10: Response Distance by Demogrpahic for 15 ambulances

```
Populations:  White: 2754 Black: 1737 Hispanic: 509
Total time passed:  16 days 13:39:49

white:  median:  1.212 mean:  1.24
black:  median:  1.054 mean:  1.106
hispanic:  median:  1.155 mean:  1.186
```

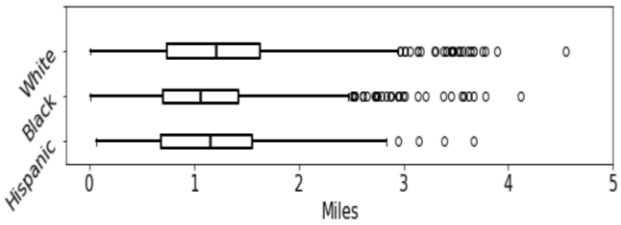

Fig. 11: Response Distance by Demogrpahic for Controll.

The distributions for white, black, and hispanic have similar traits; however, the disparate impact based for the labels previously described are not entirely inferable. Hence, tables for confidence intervals are provided in tables I through IV.
In each table, the row is the "favored class" in DI and the columns are the "unfavored class" with its 95% confidence interval.

TABLE I
DI CONFIDENCE INTERVALS FOR 25 AMBULANCES

Favored Class		
White	Black: (1.029, 1.120)	Hispanic: (0.885, 1.057)
Black	Hispanic: (0.885, 0.976)	White: (0.813, 0.994)
Hispanic	Black: (1.016, 1.198)	White: (0.944, 1.166)

TABLE II
DI CONFIDENCE INTERVALS FOR 35 AMBULANCES

Favored Class		
White	Black: (1.02, 1.112)	Hispanic: (0.933, 1.103)
Black	Hispanic: (0.892, 0.984)	White: (0.866, 1.046)
Hispanic	Black: (0.957, 1.136)	White: (0.897, 1.067)

TABLE III
DI CONFIDENCE INTERVALS FOR 15 AMBULANCES

Favored Class		
White	Black: (1.043, 1.133)	Hispanic: (0.973, 1.148)
Black	Hispanic: (0.874, 0.964)	White: (0.883, 1.066)
Hispanic	Black: (0.934, 1.118)	White: (0.855, 1.031)

TABLE IV
DI CONFIDENCE INTERVALS FOR CONTROL SIMULATION

Favored Class		
White	Black: (1.188, 1.279)	Hispanic: (0.997, 1.169)
Black	Hispanic: (0.766, 0.857)	White: (0.788, 0.969)
Hispanic	Black: (1.048, 1.229)	White: (0.837, 1.009)

The confidence intervals in the previous four tables fail to show any statistically significant bias. Moreover, RaspBary fails to show more bias than the control simulation in figures 9 and table IV.

IV. Conclusions

The simulated trails of RaspBary indicate little concern for potential algorithmic bias. Moreover, the current system of ambulances responding to emergencies from the nearest EMS station fails to show less bias than a full usage of RaspBary. In fact, the main reproducible pattern among these simulations is a slight favoring of black people. However, this could simply be embedded in the data.

According to the United States Census Bureau Marion County has the following demographic percentages: white non-hispanic at 55.5% , black or African American at 28.6%, and Hispanic or Latino at 10.5%. In the 4 simulated trails, the following average demographic percentages occurred: white non-hispanic at 55.1% , black or African American at 34.8%, and Hispanic or Latino at 10.1%.

Only African Americans had a noticeably larger proportion of emergencies relative to the Marion County distribution of demographics. If this is the true to real EMS data, then it would make sense for intensity grids to provide larger intensities in areas with more African Americans. Since EMS stations were likely created over a heatmap, a human may favor putting stations in areas of higher intensities than machine generated Wasserstein barycenters. If so, that would explain why the control simulation had a slightly higher bias favoring black people. Regardless, this is speculation and not empirically defined.

While not entirely necessary, more parity could be forced into RaspBary. Demonstrated in [9], a penalized likelihood method for predictive policing added more fairness at the cost of accuracy. A similar approach could apply to the gridded intensities generated by RaspBary. The Wasserstein barycenters of this data could be less biased. Moreover, a constraint to the algorithm in [4] could be added to ensure a minimal disparate impact regarding Wasserstein barycenters. There is not an immediate need to implement these methods, however, they are worth looking into if RaspBary had statistically significant bias for another dataset.

As far as EMS is concerned, RaspBary is safe for deployment. None of the disparate impact confidence intervals showed any acceptance of the tail end alternative hypothesis $H_a < \tau$. These trials did not demonstrate any statistically significant evidence for bias in RaspBary. These results are welcomed since refactoring RaspBary to reduce bias is not currently needed.

ACKNOWLEDGMENT

This development of RaspBary received suport from NSF grant SCC-1737585, NSF grant ATD-1737996 and NSF grant REU-1343123.

REFERENCES

[1] Besse, P., del Barrio, E., Gordaliza, P., & Loubes, J. M. (2018). Confidence intervals for testing disparate impact in fair learning. Retrieved from: https://arxiv.org/pdf/1807.06362.pdf

[2] Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and Public Policy*, *5*(1), 1-6.

[3] Bürger, A., Wnent, J., Bohn, A., Jantzen, T., Brenner, S., Lefering, R., Seewald, S., Gräsner, J., T., & Fischer, M. (2018). The effect of ambulance response time on survival following out-of-hospital cardiac arrest: an analysis from the german resuscitation registry. *Deutsches Ärzteblatt International*, *115*(33-34), 541. Retrieved from: http://dx.doi.org.proxy.ulib.uits.iu.edu/10.3238/arztebl.2018.0541

[4] Cuturi, M., and Doucet, A. (2014). Fast computation of wasserstein barycenters. *Proceedings of Machine Learning Research, 32*, 685–693.

[5] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). ACM.

[6] Hosler, R., Liu, X., Carter, J., Ganci, A., Hill, J., Raje, R., Mohler G., and Saper, M. (2018). RaspBary: hawkes point process wasserstein barycenters as a service.

[7] Hosler, R., Saper M. 2018. Hawkes process wasserstein barycenter github repository. https://github.com/rjhosler/IUPUI-REU

[8] Mohler, G., and Brantingham, P. J. (2018). Privacy preserving, crowd sourced crime hawkes processes. *2018 International Workshop on Social Sensing (SocialSens)*, 14–19. **DOI** 10.1109/SocialSens.2018.00016

[9] Mohler, G., Raje, R., Carter, J., Valasik, M., and Brantingham, P.J. (2018). A penalized likelihood method for balancing accuracy and fairness in predictive policing *IEEE International Conference on Systems, Man, and Cybernetics (SMC2018).*

[10] Reinhart, A. 2017. A review of self-exciting spatio-temporal point process and their applications. *Statistical Science.*