MORPH-II Dataset

Summary and Cleaning

Garrett Bingham & Ben Yip

June 16, 2017

University of North Carolina Wilmington

Table of contents

Introduction to the Data

MORPH-II: An Overview

The MORPH-II dataset is composed of mugshots of people from 16 to 77 years of age, with an average of 4 images per person. It is the largest longitudinal face image dataset publicly available. The academic version (which we use) contains roughly 55,000 images taken over 5 years, while the commercial version has about 202,000 images spanning 8 years.

See: `https://ebill.uncw.edu/C20231_ustores/web/store_main.jsp?STOREID=4`

MORPH-II: Metadata

morph_2008_nonCommercial.csv

This release contains 11 variables:

id_num

6-digit subject identifier

picture_num

subject photo number

dob

date of birth (mm/dd/yyyy)

doa

date of arrest (mm/dd/yyyy)

race

(B, W, A, H, O)

gender

(M or F)

facial_hair

not recorded (NULL)

age

integer age ($\lfloor doa - dob \rfloor$)

age_diff

time since last arrest (days)

glasses

not recorded (NULL)

photo

image filename

MORPH-II: Demographic Makeup

The MORPH-II dataset is a collection of 55,134 mugshots, including many of repeat offenders (providing valuable longitudinal data). The below table summarizes the demographic composition of the dataset.

Table 1: Number of Images by Gender and Race

	Black	White	Asian	Hispanic	Other	**Total**
Male	36,832	7,961	141	1,667	44	46,645
Female	5,757	2,598	13	102	19	8,489
Total	42,589	10,559	154	1,769	63	**55,134**

Note:

This table was taken from the original MORPH Non-Commercial Release Whitepaper. After cleaning, the total number of images is the same but individual values may be slightly different.

MORPH-II: Summary Info

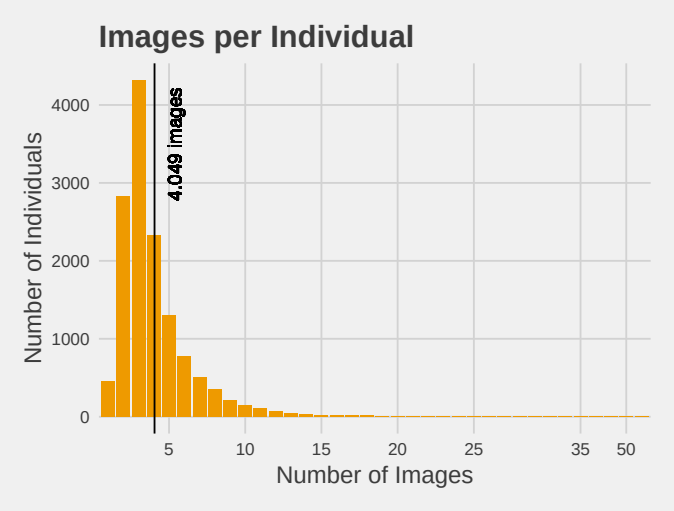


Figure 1: Barplot of Images per Subject

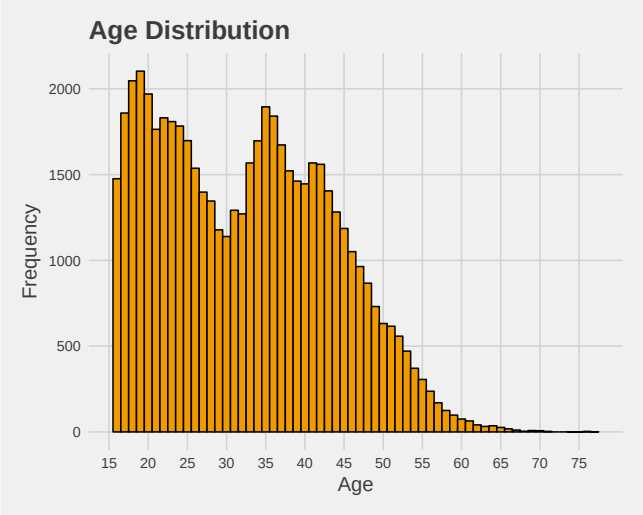


Figure 2: MORPH-II Age Distribution

Table 2: Number of Distinct Individuals

	Distinct Individuals
Male	11,459
Female	2159
Sum vs. Total	13,618 vs. 13,617

Inconsistencies in the Data

Inconsistencies in the Data

Repeat offenders have multiple entries in the MORPH-II dataset. There are some people with more than one gender, race, and/or birthdate. This causes problems when trying to use the images to predict demographics.

Table 3: MORPH-II Inconsistencies by Attribute

Attribute	Number of People
Gender	1
Race	33
Birthdate	1779

Cleaning the Data

Cleaning the Data: Gender

(a) Female

(b) Male

(c) Female

(d) Female

(e) Female

(f) Female

Cleaning the Data: Race

(1a) White

(1b) Black

(1c) White

(2a) Asian

(2b) White

(2c) Black

Person 1 has 24 images classified as White and 1 image classified as Black

Cleaning the Data: Race

Each of the 33 people with inconsistent race was evaluated on a case by case basis. A final decision was made according to one of the following criteria:

Simple Majority

All images for a given person were assigned the race that appeared at least 50% of the time.

Visual Estimation

Each person's images were inspected one at a time. We decided the race only if there was a wide consensus among our team members.

Other

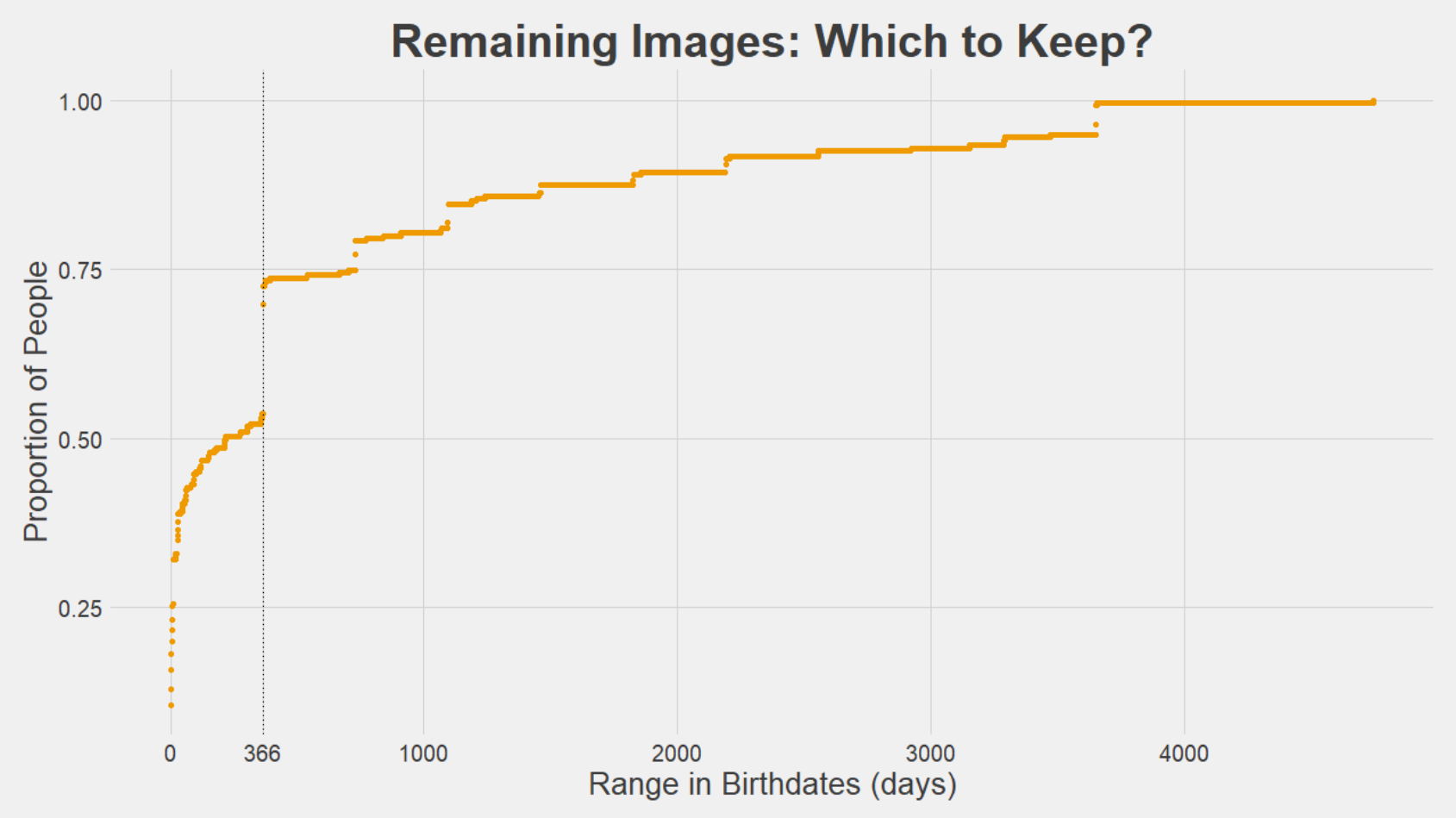
For some people (e.g. those of mixed race) it was difficult to guess their race from the photos, and there was substantial variation in the original dataset. We set the race of all images to **Other**.

Cleaning the Data: Birthdate

Similar to cleaning the race data, we were able to use a simple majority for 1524 of the 1779 people with inconsistent birthdates.

However, the remaining 255 people posed additional problems. For some of them, their birthdates were in a multiway tie. For others, there was no majority, or their birthdates differed by several years. This made it difficult to choose one birthdate over another.

Cleaning the Data: Birthdate



For each person whose birthdates differed by no more than one year, we calculated the mean birthdate and assigned this date to all images. The remaining images were set aside as *Not For Training*.

Cleaning the Data: Birthdate

Table 4: Cleaned Data Summary

Solution	Number of People	Number of Images
Simple Majority	1524	1906
Average Birthdate	185	515
Not For Training	70	230
Total	**1779**	**2651**

New Datasets

New Datasets

After being cleaned, the data was divided into 3 new files:

morphII_cleaned_v2

This file is the same as morph_2008_nonCommercial.csv, but with dob, race, and gender inconsistencies corrected.

morphII_go_for_age

Individuals with incorrectable birthdates were removed from the above dataset. This leaves all the images with consistent age information that are ready for training and testing age estimation models.

morphII_holdout_for_age

These are the images (mentioned above) with incorrectable birthdates.

New Variables

Each of the new datasets also has two additional variables:

corrected
indicator (0-8)

age_dec
decimal age (*doa* − *dob*)

About corrected

The corrected column contains an indicator variable which takes a different value depending on whether or not it was modified. Unchanged observations are labeled as 0, while those that were corrected or marked for hold out take a value between 1 and 8 depending on what was done to them.

New Datasets: Updated Info

Table 5: Cleaned Data - Number of Images by Gender and Race

	Black	White	Asian	Hispanic	Other	Total
Male	36,821	7,958	140	1,661	64	46,644
Female	5,756	2,590	13	99	32	8,490
Total	42,577	10,548	153	1,760	96	**55,134**

Table 6: Net Change in Number of Images by Gender and Race

	Black	White	Asian	Hispanic	Other	Total
Male	-11	-3	-1	-6	+20	-1
Female	-1	-8	-0	-3	+13	+1
Total	-12	-11	-1	-9	+33	**-0**

New Datasets: Updated Info

Table 7: Original Data - Number of Distinct Individuals

	Black	White	Asian	Hispanic	Other	Total
Male	8838	2070	49	517	15	11489
Female	1494	634	6	30	5	2169
Total	10332	2704	55	547	20	13658

Table 8: Cleaned Data - Number of Distinct Individuals

	Black	White	Asian	Hispanic	Other	Total
Male	8829	2056	47	507	19	11458
Female	1491	628	4	28	8	2159
Total	10320	2684	51	535	27	**13617**

Dirty Data

Dirty Data: Examples of Research on Uncleaned MORPH-II

G. Guo and G. Mu.
Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression.
In *CVPR 2011*, pages 657–664, June 2011.

K. H. Liu, S. Yan, and C. C. J. Kuo.
Age estimation via grouping and decision fusion.
IEEE Transactions on Information Forensics and Security, 10(11):2408–2423, Nov 2015.

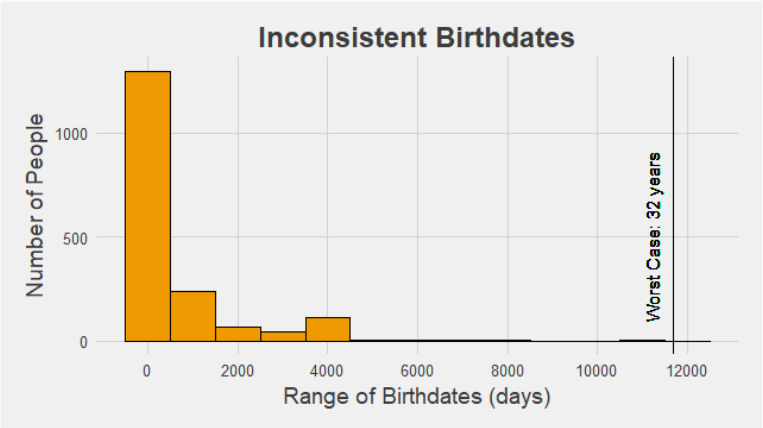
X. Wang, V. Ly, G. Lu, and C. Kambhamettu.
Can we minimize the influence due to gender and race in age estimation?
In *2013 12th International Conference on Machine Learning and Applications*, volume 2, pages 309–314, Dec 2013.

D. H. P. Yassin, S. Hoque, and F. Deravi.
Age sensitivity of face recognition algorithms.
In *2013 Fourth International Conference on Emerging Security Technologies*, pages 12–15, Sept 2013.

This is merely a sampling. Many other articles exist that used an uncleaned version of MORPH-II.

Dirty Data: Consequences of Using Uncleaned MORPH-II

There will not likely be an enormous impact on model performance for gender or race prediction, because the number of gender and race inconsistencies is small.



Age estimation models will see a drop in overall performance manifest in a higher Mean Absolute Error (MAE). For some people in the dataset, their birthdates vary enough that their *age decreases with time*. This will significantly affect models concerned with age progression.

Conclusion

Conclusion: Clean Data Matters

Cleaning the data before doing research is vital. This not only preserves the accuracy of one's results, but also the integrity. Many researchers base their work off of previous results, making it even more important to ensure that one's own work is accurate.