BeeBDC Duplicate Status by Data Source Analysis

Gretchen LeBuhn

2025-06-03

Table of contents

1 Overview

This analysis examines the distribution of duplicate status classifications across different data sources in the Alarcon-Cruz *et al.* dataset.

```
# Load data
data <- read.csv("~/GitHub/phylo-endemism/data/BeeDataNoiNat_Clean.csv", stringsAsFactors = FA

cat("Total records in dataset:", format(nrow(data), big.mark = ","), "\n")
```

```
Total records in dataset: 485,231
```

```
cat("Records with source information:", format(sum(!is.na(data$source) & data$source != ""), bi
```

```
Records with source information: 485,231
```

```
cat("Records with duplicate status:", format(sum(!is.na(data$duplicateStatus) & data$duplicateS
```

```
Records with duplicate status: 485,231
```

2 Overall Duplicate Status Distribution

```
# Overall duplicate status summary
overall_duplicate_summary <- data %>%
  filter(!is.na(duplicateStatus) & duplicateStatus != "") %>%
  count(duplicateStatus, sort = TRUE) %>%
  mutate(percentage = round(n / sum(n) * 100, 1))

kable(overall_duplicate_summary,
      caption = "Overall Distribution of Duplicate Status Classifications",
      col.names = c("Duplicate Status", "Records", "Percentage %"),
      format.args = list(big.mark = ","))
```

Table 1: Overall Distribution of Duplicate Status Classifications

Duplicate Status	Records	Percentage %
Unique	278,747	57.4
Kept duplicate	206,484	42.6

3 Data Sources Summary

```
# Summary of data sources - unique specimens only
source_summary <- data %>%
  filter(!is.na(source) & source != "" &
         (str_detect(tolower(duplicateStatus), "kept") |
          str_detect(tolower(duplicateStatus), "unique") |
          duplicateStatus == "" |
          is.na(duplicateStatus))) %>%
  count(source, sort = TRUE) %>%
  mutate(percentage = round(n / sum(n) * 100, 1)) %>%
  rename(unique_specimens = n)

cat("Number of data sources:", nrow(source_summary), "\n")
```

```
Number of data sources: 3
```

```r
cat("Total  specimens with source information:", format(sum(source_summary$unique_specimens), b
```

```
Total  specimens with source information: 485,231
```

```r
kable(source_summary,
      caption = "Specimens by Data Source",
      col.names = c("Data Source", "Unique Specimens", "Percentage %"),
      format.args = list(big.mark = ","))
```

Table 2: Specimens by Data Source

Data Source	Unique Specimens	Percentage %
Big Bee	222,324	45.8
BeeBDC	142,391	29.3
HIkerd	120,516	24.8

4 Duplicate Status by Data Source

```r
# Create cross-tabulation of source vs duplicate status
duplicate_by_source <- data %>%
  filter(!is.na(source) & source != "" &
         !is.na(duplicateStatus) & duplicateStatus != "") %>%
  count(source, duplicateStatus) %>%
  pivot_wider(names_from = duplicateStatus, values_from = n, values_fill = 0)

# Add row totals
duplicate_by_source <- duplicate_by_source %>%
  mutate(Total = rowSums(select(., -source))) %>%
  arrange(desc(Total))

# Display as table
kable(duplicate_by_source,
      caption = "Duplicate Status Counts by Data Source",
      format.args = list(big.mark = ","))
```

Table 3: Duplicate Status Counts by Data Source

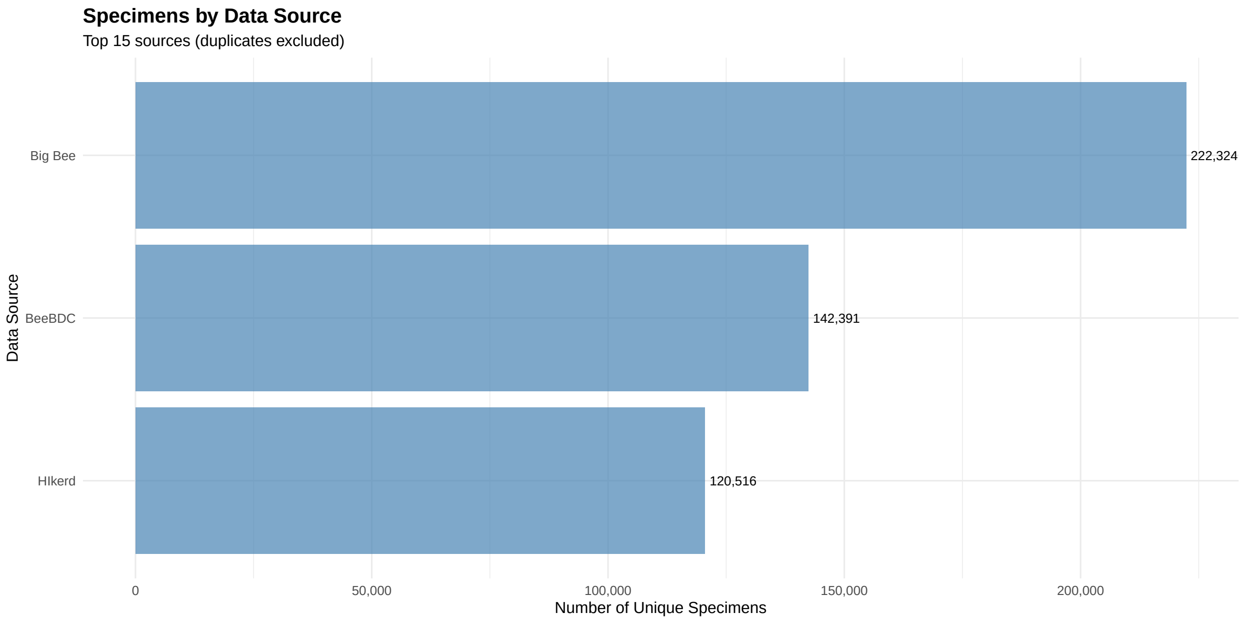
source	Kept duplicate	Unique	Total
Big Bee	201,689	20,635	222,324
BeeBDC	294	142,097	142,391
HIkerd	4,501	116,015	120,516

5 Visualization: Unique Specimens by Data Source

```r
# Get top sources for visualization
top_sources_for_plot <- source_summary %>%
  head(15) %>%
  pull(source)

# Create bar chart for unique specimens by source
plot_data <- source_summary %>%
  head(15) %>%
  mutate(source = reorder(source, unique_specimens))

ggplot(plot_data, aes(x = source, y = unique_specimens)) +
  geom_col(fill = "steelblue", alpha = 0.7) +
  geom_text(aes(label = format(unique_specimens, big.mark = ",")),
            hjust = -0.1, size = 3) +
  coord_flip() +
  labs(
    title = "Specimens by Data Source",
    subtitle = "Top 15 sources (duplicates excluded)",
    x = "Data Source",
    y = "Number of Unique Specimens"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold")
  ) +
  scale_y_continuous(labels = scales::comma)
```



6 Institutional Analysis - Only specimen data used in analysis

```r
# Function to extract institution codes from catalog numbers
extract_institution <- function(catalog_num) {
  if(is.na(catalog_num) || catalog_num == "") return(NA)

  # Extract alphabetic prefix before numbers
  prefix <- str_extract(catalog_num, "^[A-Za-z]+")
  return(prefix)
}

# Filter to only unique specimens (exclude duplicates)
unique_specimens <- data %>%
  filter(str_detect(tolower(duplicateStatus), "kept") |
         str_detect(tolower(duplicateStatus), "unique") |
         duplicateStatus == "" |
         is.na(duplicateStatus))

# Apply institution extraction to unique specimens only
unique_specimens$institution_code <- sapply(unique_specimens$catalogNumber, extract_institutio

# Create institution lookup with common codes
institution_lookup <- data.frame(
  code = c("AMNH", "ANSP", "BBSL", "BMNH", "CAS", "EMEC", "FSCA", "INHS",
           "KUNHM", "LACM", "MCZ", "MSUC", "NMNH", "OSUC", "PCYU", "SDNHM",
           "SEMC", "TAMU", "UAIC", "UBCZ", "UCB", "UCDC", "UCMS", "UCRC",
           "UCD", "USNM", "WIRC", "YPM"),
  institution = c("American Museum of Natural History",
                  "Academy of Natural Sciences of Philadelphia",
                  "Bee Biology and Systematics Laboratory",
                  "Natural History Museum, London",
                  "California Academy of Sciences",
                  "Essig Museum of Entomology, UC Berkeley",
                  "Florida State Collection of Arthropods",
                  "Illinois Natural History Survey",
                  "Kansas University Natural History Museum",
                  "Los Angeles County Museum",
                  "Museum of Comparative Zoology, Harvard",
                  "Michigan State University",
                  "National Museum of Natural History",
                  "Ohio State University Collection",
                  "Pacific Coast Entomological Society",
                  "San Diego Natural History Museum",
                  "Snow Entomological Museum, Kansas",
                  "Texas A&M University",
                  "University of Arizona Insect Collection",
                  "University of British Columbia",
                  "University of California, Berkeley",
                  "UC Davis Center for Population Biology",
                  "UC Museum of Paleontology",
                  "UC Riverside Entomology Collection",
                  "UC Davis Entomology Collection",
                  "US National Museum",
                  "Wisconsin Insect Research Collection",
                  "Yale Peabody Museum"),
  stringsAsFactors = FALSE
)

# Count unique specimens by institution code
institution_summary_unique <- unique_specimens %>%
  filter(!is.na(institution_code) & institution_code != "") %>%
  count(institution_code, name = "Unique_Specimens") %>%
  left_join(institution_lookup, by = c("institution_code" = "code")) %>%
  mutate(institution = ifelse(is.na(institution),
                             institution_code,  # Use the code itself instead of "Unknown Insti
                             institution)) %>%
  arrange(desc(Unique_Specimens))

cat("Total institutions identified from unique specimens:", nrow(institution_summary_unique), "
```

```
Total institutions identified from unique specimens: 171
```

```
cat("Unique specimens with institution codes:",
    format(sum(institution_summary_unique$Unique_Specimens), big.mark = ","), "\n")
```

```
Unique specimens with institution codes: 419,340
```

```
cat("Total unique specimens in dataset:",
    format(nrow(unique_specimens), big.mark = ","), "\n")
```

```
Total unique specimens in dataset: 485,231
```

6.1 Institution Summary Table - Specimens used in analysis only

```
# Create comprehensive institution table for unique specimens
institution_table_unique <- institution_summary_unique %>%
  mutate(Percentage = round(Unique_Specimens / sum(Unique_Specimens) * 100, 1)) %>%
  select(institution_code, institution, Unique_Specimens, Percentage) %>%
  rename("Institution Code" = institution_code,
         "Institution Name" = institution,
         "Number of Specimens" = Unique_Specimens,
         "Percentage %" = Percentage)

# For PDF output, use kable instead of datatable
kable(institution_table_unique,
      caption = "Institutional Sources of Bee Specimens used in analysis (Duplicates Excluded)"
      format.args = list(big.mark = ","))
```

Table 4: Institutional Sources of Bee Specimens used in analysis (Duplicates Excluded)

Institution Code	Institution Name	Number of Specimens	Percentage %
BBSL	Bee Biology and Systematics Laboratory	86,758	20.7
PINN	PINN	73,106	17.4
UCRC	UC Riverside Entomology Collection	43,879	10.5
EMEC	Essig Museum of Entomology, UC Berkeley	33,078	7.9
YOSE	YOSE	21,063	5.0
SFSU	SFSU	16,953	4.0
JPS	JPS	14,992	3.6
BBSLID	BBSLID	13,891	3.3
AMNH	American Museum of Natural History	13,463	3.2
KWC	KWC	9,235	2.2
LACM	Los Angeles County Museum	8,916	2.1
USGS	USGS	6,889	1.6
BMEC	BMEC	6,703	1.6
USNMENT	USNMENT	6,586	1.6
UCSB	UCSB	6,581	1.6
UCSCRMIC	UCSCRMIC	6,087	1.5
INHS	Illinois Natural History Survey	5,547	1.3
BMEP	BMEP	5,074	1.2
UCFC	UCFC	5,047	1.2
LACMENT	LACMENT	2,994	0.7
CSCA	CSCA	2,717	0.6
M	M	2,378	0.6
PUB	PUB	2,159	0.5
UPLOAD	UPLOAD	2,133	0.5
FDP	FDP	1,510	0.4
X	X	1,508	0.4
UCIS	UCIS	1,506	0.4
UCRCENT	UCRCENT	1,470	0.4
CHIS	CHIS	1,461	0.3
DRO	DRO	1,196	0.3
SDNHM	San Diego Natural History Museum	1,155	0.3

Institution Code	Institution Name	Number of Specimens	Percentage %
UCREM	UCREM	946	0.2
YPM	Yale Peabody Museum	897	0.2
SAMO	SAMO	626	0.1
FMNHINS	FMNHINS	619	0.1
OSUC	Ohio State University Collection	618	0.1
USNM	US National Museum	603	0.1
UMMZI	UMMZI	594	0.1
FSCA	Florida State Collection of Arthropods	567	0.1
DEVA	DEVA	540	0.1
BOMBUS	BOMBUS	492	0.1
Berk	Berk	471	0.1
Morandin	Morandin	435	0.1
SBMNHENT	SBMNHENT	419	0.1
ASUHIC	ASUHIC	377	0.1
REDW	REDW	363	0.1
OS	OS	351	0.1
OSMIA	OSMIA	311	0.1
CASENT	CASENT	297	0.1
TS	TS	215	0.1
none	none	215	0.1
MOJA	MOJA	202	0.0
PORE	PORE	169	0.0
Davis	Davis	166	0.0
JBWM	JBWM	158	0.0
CUIC	CUIC	157	0.0
JOTR	JOTR	157	0.0
BLMMLP	BLMMLP	134	0.0
UCSC	UCSC	134	0.0
Step	Step	116	0.0
BIOUG	BIOUG	112	0.0
CAS	California Academy of Sciences	97	0.0
DVNM	DVNM	97	0.0
Ribb	Ribb	91	0.0
CSU	CSU	84	0.0
RGL	RGL	84	0.0
Boha	Boha	73	0.0
AMNHBEE	AMNHBEE	60	0.0
RLMC	RLMC	59	0.0
LaKU	LaKU	56	0.0
RUAC	RUAC	44	0.0
TUZ	TUZ	43	0.0
UAIC	University of Arizona Insect Collection	43	0.0
NRidg	NRidg	40	0.0
ZMA	ZMA	40	0.0
SACR	SACR	38	0.0

Institution Code	Institution Name	Number of Specimens	Percentage %
UCMC	UCMC	37	0.0
MSU	MSU	34	0.0
SanF	SanF	34	0.0
River	River	33	0.0
Timb	Timb	33	0.0
UTEP	UTEP	33	0.0
LaBer	LaBer	31	0.0
NMDG	NMDG	29	0.0
Fresno	Fresno	28	0.0
UNK	UNK	26	0.0
Mich	Mich	25	0.0
OSAC	OSAC	24	0.0
CCDB	CCDB	21	0.0
UMNH	UMNH	21	0.0
KJH	KJH	18	0.0
BT	BT	17	0.0
CASTYPE	CASTYPE	17	0.0
Pinn	Pinn	17	0.0
UCMS	UC Museum of Paleontology	17	0.0
Daly	Daly	16	0.0
Grig	Grig	16	0.0
Thorp	Thorp	16	0.0
FD	FD	15	0.0
PYU	PYU	15	0.0
USGSDRO	USGSDRO	15	0.0
BerkJHC	BerkJHC	14	0.0
OBS	OBS	14	0.0
FORB	FORB	13	0.0
LACo	LACo	13	0.0
RH	RH	13	0.0
RSKM	RSKM	13	0.0
SRFS	SRFS	13	0.0
VTEC	VTEC	13	0.0
PSUC	PSUC	11	0.0
SEMC	Snow Entomological Museum, Kansas	10	0.0
http	http	9	0.0
DIAL	DIAL	8	0.0
HOLO	HOLO	8	0.0
HYM	HYM	8	0.0
NMSUACP	NMSUACP	8	0.0
NCSU	NCSU	7	0.0
UNHC	UNHC	7	0.0
WRME	WRME	6	0.0
Zavo	Zavo	6	0.0
mojave	mojave	6	0.0

Institution Code	Institution Name	Number of Specimens	Percentage %
personal	personal	6	0.0
Gain	Gain	5	0.0
NAUF	NAUF	5	0.0
RMNH	RMNH	5	0.0
WFBM	WFBM	5	0.0
Wash	Wash	5	0.0
casent	casent	5	0.0
D	D	4	0.0
NewY	NewY	4	0.0
Sacr	Sacr	4	0.0
MEM	MEM	3	0.0
Moscow	Moscow	3	0.0
Osmia	Osmia	3	0.0
TAMZ	TAMZ	3	0.0
BBLMMLP	BBLMMLP	2	0.0
CMNHENT	CMNHENT	2	0.0
Colo	Colo	2	0.0
GEB	GEB	2	0.0
GMP	GMP	2	0.0
Gris	Gris	2	0.0
LMNRA	LMNRA	2	0.0
Lincoln	Lincoln	2	0.0
MSUC	Michigan State University	2	0.0
Park	Park	2	0.0
bbsl	bbsl	2	0.0
ASUHIV	ASUHIV	1	0.0
BBSl	BBSl	1	0.0
BLCU	BLCU	1	0.0
BREM	BREM	1	0.0
BSSL	BSSL	1	0.0
Cornel	Cornel	1	0.0
Donov	Donov	1	0.0
ERRR	ERRR	1	0.0
FOBU	FOBU	1	0.0
Hurd	Hurd	1	0.0
IZBE	IZBE	1	0.0
LA	LA	1	0.0
Lins	Lins	1	0.0
PCYU	Pacific Coast Entomological Society	1	0.0
RHS	RHS	1	0.0
ROM	ROM	1	0.0
SAM	SAM	1	0.0
SDC	SDC	1	0.0
ST	ST	1	0.0
SanJ	SanJ	1	0.0

Institution Code	Institution Name	Number of Specimens	Percentage %
TTU	TTU	1	0.0
UCR	UCR	1	0.0
USCB	USCB	1	0.0
WATR	WATR	1	0.0
bbslid	bbslid	1	0.0

```r
# Show top 20 for summary
cat("\n\nTop 20 Institutions by Number of Bee Specimens used in analysis (Unique):\n")
```

```
Top 20 Institutions by Number of Bee Specimens used in analysis (Unique):
```

```r
kable(head(institution_table_unique, 20),
      caption = "Top 20 Institutions by Number of Bee Specimens used in analysis (Unique)",
      format.args = list(big.mark = ","))
```

Table 5: Top 20 Institutions by Number of Bee Specimens used in analysis (Unique)

Institution Code	Institution Name	Number of Specimens	Percentage %
BBSL	Bee Biology and Systematics Laboratory	86,758	20.7
PINN	PINN	73,106	17.4
UCRC	UC Riverside Entomology Collection	43,879	10.5
EMEC	Essig Museum of Entomology, UC Berkeley	33,078	7.9
YOSE	YOSE	21,063	5.0
SFSU	SFSU	16,953	4.0
JPS	JPS	14,992	3.6
BBSLID	BBSLID	13,891	3.3
AMNH	American Museum of Natural History	13,463	3.2
KWC	KWC	9,235	2.2
LACM	Los Angeles County Museum	8,916	2.1
USGS	USGS	6,889	1.6
BMEC	BMEC	6,703	1.6
USNMENT	USNMENT	6,586	1.6
UCSB	UCSB	6,581	1.6
UCSCRMIC	UCSCRMIC	6,087	1.5
INHS	Illinois Natural History Survey	5,547	1.3
BMEP	BMEP	5,074	1.2
UCFC	UCFC	5,047	1.2

Institution Code	Institution Name	Number of Specimens	Percentage %
LACMENT	LACMENT	2,994	0.7