KDE Analysis Summary & Workflow Guide

Updated: December 30, 2024

FINAL CONFIGURATION

Data Filters

- **Date range**: 2015-01-01 to 2024-12-31 (10 years)
- **Species threshold**: ≥25 records for individual species KDEs
- **Taxa**: Cetaceans only (odontocetes and mysticetes)

Bandwidth Settings (sigma in meters)

```r
DEFAULT_SIGHTINGS_SIGMA <- 10000  # Most species
DEFAULT_PAM_SIGMA <- 5000         # All PAM species

SPECIES_SPECIFIC_SIGMA <- list(
  "WHALE-NORTHERN BOTTLENOSE" = 5000,  # Reduced from 10km
  "WHALE-SOWERBY'S BEAKED" = 5000      # Reduced from 10km
)
```

Percentile Thresholding

```r
SIGHTINGS_PERCENTILE_THRESHOLD <- 0.50  # Remove bottom 50%
PAM_PERCENTILE_THRESHOLD <- 0.25        # Remove bottom 25%
QUANTILE_PROBS <- c(0.85, 0.9, 0.95, 1) # High-use areas only
```

WORKFLOW SCRIPTS

1. `run_all_kdes_fixed.R` - KDE Generation

Purpose: Generate KDE rasters and standard quantile shapefiles

What it does:

1. Cleans output folders (removes old .tif and .png files)
2. Loads sightings data with 2015+ filter
3. Creates KDE rasters with species-specific bandwidths
4. Saves rasters to `output/tif/`
5. Generates diagnostic plots to `output/FIGS/`
6. **NEW**: Converts all rasters to standard quantile shapefiles (50%, 85%, 90%, 95%)
7. Saves standard shapefiles to `output/shapes/`

Key features:

- Uses common study area window for all species
- Applies species-specific bandwidths for NBW and Sowerby's
- Normalized KDE values (0-1 scale)
- Creates both individual species and family-level KDEs

Outputs:

- `output/tif/*.tif` - KDE rasters
- `output/shapes/*.shp` - Standard quantile shapefiles
- `output/FIGS/*.png` - Diagnostic plots

---

2. `kde_threshold_workflow_fixed.R` - Threshold Processing & Comparison

Purpose: Apply percentile thresholding and create comparison maps

What it does:

1. Reads KDE rasters from `output/tif/`
2. Applies percentile threshold (50% for sightings, 25% for PAM)
3. Calculates quantiles on filtered data (85%, 90%, 95%)
4. Creates threshold shapefiles in `output/shapes/percentile_threshold/`
5. Loads standard shapefiles from `output/shapes/`
6. Creates side-by-side comparison maps
7. Filters 50% quantile from standard for visual comparison
8. Uses unified color palette across both maps

Key fixes implemented:

- Species name matching handles spaces, underscores, and hyphens
- Consistent 2015+ date filtering for point overlays
- Separate thresholds for PAM vs sightings
- Proper filename extraction for all species patterns

Outputs:

- `output/shapes/percentile_threshold/*.shp` - Threshold shapefiles
- `output/FIGS/QA_threshold_comparison/*.png` - Comparison maps

KEY DECISIONS & RATIONALE

1. Why 2015+ Date Filter?

- Ensures temporal consistency with recent survey effort
- Excludes older data with different survey protocols
- 10-year window provides sufficient sample size
- **CRITICAL**: Both KDE generation AND comparison plotting use same date filter

2. Why Different PAM vs Sightings Thresholds?

Sightings (50%):

- Wider spatial extent from bandwidth extrapolation
- More prone to "donut" artifacts from single points
- Needs aggressive filtering to focus on core areas

PAM (25%):

- Already spatially constrained by station locations
- Less prone to edge effects
- Lower threshold preserves more spatial detail

3. Why Species-Specific Bandwidths for NBW and Sowerby's?

- Very sparse, highly concentrated distributions
- Standard 10km bandwidth creates excessive smoothing
- 5km bandwidth better captures actual distribution
- Reduces isolated donut artifacts

4. Why Remove 50% Quantile from Standard Maps?

- Visual comparison issue: 50% in standard looked like 85% in threshold
- Different number of quantiles = different color assignments
- Removing 50% from standard creates fair comparison
- Focus on high-use areas (85%+) in both approaches

METHODOLOGICAL ISSUES & SOLUTIONS

Issue 1: "Donut Problem"

Problem: Single or few sightings create circular high-density artifacts that rank in top quantiles

Root cause:

- Gaussian kernel creates smooth density surface around each point
- Isolated points produce circular contours
- With few observations, these rank high in quantile classification

Solutions implemented:

- Percentile thresholding removes bottom 50% of density values
- Species-specific smaller bandwidths for sparse species
- Combined approach addresses both smoothing and classification issues

Issue 2: Filename Matching Chaos

Problem: Multiple naming patterns across files made matching difficult

- Standard: `KDE_combined_species_WHALE-NORTHERN_BOTTLENOSE.shp`
- Threshold: `KDE_WHALE-NORTHERN_BOTTLENOSE_pct50.shp`
- Search: Looking for `WHALE-NORTHERNBOTTLENOSE`

Solution: Normalize both search pattern AND filenames by removing spaces and underscores

```
r

species_name_clean <- toupper(gsub("[ _]", "", species_name))
fname_upper <- toupper(gsub("[ _]", "", basename(f)))
```

Issue 3: Date Filter Mismatch

Problem: KDEs generated from 2010+ data but comparison points filtered to 2015+

- Created ghost contours with no points inside
- Particularly visible for species with pre-2015 concentration

Solution: Applied consistent 2015+ filter in BOTH scripts

- KDE generation filters at data loading
- Comparison script filters before plotting points

Issue 4: Standard Shapefiles Not Updating

Problem: Running KDE script but comparison still showed old data

- Rasters were updated but shapefiles weren't
- Scripts were separate, easy to forget second step

Solution: Integrated shapefile generation into KDE script

- Now happens automatically after raster creation
- One script, one run, everything synced

FILE STRUCTURE

```
output/
├── tif/                          # KDE rasters (normalized 0-1)
│   ├── KDE_combined_species_*.tif
│   ├── KDE_pam_*.tif
│   ├── KDE_combined_odontocetes_*.tif
│   └── KDE_combined_mysticetes_*.tif
│
├── shapes/                       # Standard quantile shapefiles
│   ├── KDE_combined_species_*.shp    # 50%, 85%, 90%, 95%
│   └── KDE_pam_*.shp
│
├── shapes/percentile_threshold/      # Threshold quantile shapefiles
```

```
│   ├── KDE_*_pct50.shp          # Sightings (85%, 90%, 95% of top 50%)
│   └── KDE_*_pct25.shp          # PAM (85%, 90%, 95% of top 75%)
│
└── FIGS/
    ├── *.png                    # Individual diagnostic plots
    └── QA_threshold_comparison/     # Side-by-side comparisons
        ├── compare_sightings_*.png
        └── compare_pam_*.png
```

RUNNING THE ANALYSIS

Step 1: Generate KDEs

```
r

# In run_all_kdes_fixed.R, set:
RUN_SIGHTINGS <- TRUE
RUN_PAM <- FALSE  # or TRUE if processing PAM

# Run the script
source("run_all_kdes_fixed.R")
```

What to check:

- Console shows "2015+" date filter applied
- Species counts match expectations (≥25 records)
- All species processed without errors
- Shapefiles created in `output/shapes/`

Step 2: Generate Thresholds & Comparisons

```
r

source("kde_threshold_workflow_fixed.R")
```

What to check:

- Threshold shapefiles created for all species
- Comparison maps show points overlapping with KDEs (no ghost contours)
- Color consistency between standard and threshold maps
- NBW and Sowerby's show reduced donut artifacts

QUALITY CONTROL CHECKLIST

Before Running Analysis

- ☐ Input data file exists and is up to date
- ☐ Study area shapefile exists
- ☐ Output folders exist (will be created if not)

After KDE Generation

- ☐ Check date range in console output (should be 2015-2024)
- ☐ Check species counts (should be ≥25 for processed species)
- ☐ Verify raster files in `output/tif/` have today's date
- ☐ Verify shapefile files in `output/shapes/` have today's date
- ☐ Review diagnostic plots for obvious issues

After Threshold Comparison

- ☐ All species have comparison maps created
- ☐ No "Shapefiles not found" errors in console
- ☐ Points in comparison maps overlap with KDE contours
- ☐ No isolated donut contours without points
- ☐ NBW and Sowerby's show improvement over standard
- ☐ Color legends match between standard and threshold

Red Flags to Watch For

⚠️ **Ghost contours**: Contours with no points = date filter mismatch ⚠️ **Missing threshold files**: Check

filename extraction worked ⚠️ **All data filtered out**: Threshold too aggressive for that species ⚠️ **Excessive donuts**: Bandwidth may need adjustment

---

TROUBLESHOOTING

"Shapefiles not found" error

Cause: Filename matching failed **Fix**: Check console output for "Looking for pattern" - verify files exist with similar names

Points don't match KDE contours

Cause: Date filter mismatch between scripts **Fix**: Verify both scripts filter to 2015-01-01

Species has no threshold shapefile

Cause: All data filtered out by threshold **Fix**: Lower the threshold for that species or check if it has sufficient data

NBW still has donuts

Options:

1. Reduce bandwidth further (try 3000m)
2. Increase threshold to 75%
3. Both

Script runs but shapefiles not created

Cause: Script stopped/crashed before shapefile generation **Fix**: Let script run to completion, check for errors

---

METHODS TEXT FOR PAPER

Kernel Density Estimation

> "Kernel density estimation was performed using a Gaussian kernel with bandwidths of 10 km for most species and 5 km for species with highly concentrated distributions (Northern Bottlenose Whale and Sowerby's Beaked Whale). For passive acoustic monitoring (PAM) data, a 5 km bandwidth was used to reduce spatial smoothing artifacts from station-based sampling. All analyses used data from 2015-2024 to ensure temporal consistency with recent survey efforts."

Percentile Thresholding

> "To focus on core-use areas and minimize edge effects from bandwidth extrapolation, we filtered cells below the 50th percentile of kernel density for sightings data and the 25th percentile for PAM data before classifying remaining areas into high-use quantiles (85%, 90%, 95%). The different thresholds reflect the inherent spatial characteristics of each data type: sightings data require more aggressive filtering due to wider spatial extent from bandwidth effects, while PAM data are inherently spatially constrained by station locations."

Quantile Classification

> "High-use areas were classified using quantile-based thresholds (85%, 90%, 95%) calculated on the filtered density surfaces. This approach ensures that quantile classifications reflect true concentration areas rather than including low-density peripheral zones from kernel bandwidth effects."

OUTSTANDING ITEMS / FUTURE CONSIDERATIONS

If NBW/Sowerby's Still Problematic

1. **Option A**: Further reduce bandwidth to 3000m
2. **Option B**: Implement species-specific thresholds (75% for sparse species)
3. **Option C**: Use minimum sample size filter (requires point count raster)

If PAM Still Shows Bullseye

1. **Regenerate PAM KDEs** with smaller bandwidth (3000-4000m) at source
2. Adjust PAM threshold (try 10% or 15%)
3. Consider using only 90% and 95% quantiles for PAM

Alternative Approaches Discussed But Not Implemented

1. **Minimum sample size filter**: Only include areas with X+ observations
    - Most defensible methodologically
    - Requires additional spatial join to count points per cell
2. **Cluster-based filtering**: Remove small isolated patches
    - Could use area threshold on polygons
    - Risk of removing legitimate small hotspots
3. **Just use higher quantiles**: Skip 50%, 75%, only show 85%+
    - Simpler, more transparent
    - Doesn't address underlying donut problem

SCRIPT MAINTENANCE NOTES

Critical Code Sections

1. **Date filtering** (line ~337 in run_all_kdes): Must match comparison script
2. **Species name extraction** (line ~130-156 in threshold): Handles all naming patterns
3. **Filename matching** (line ~337-356 in threshold): Normalizes spaces/underscores
4. **Threshold detection** (line ~68-73 in threshold): PAM vs sightings logic

Common Modifications

- **Change date range**: Update `filter(as.Date(date_utc) >= as.Date("YYYY-MM-DD"))`
- **Change species threshold**: Update `filter(n_records >= XX)`
- **Change bandwidths**: Update `SPECIES_SPECIFIC_SIGMA` list
- **Change thresholds**: Update `PAM_PERCENTILE_THRESHOLD` and `SIGHTINGS_PERCENTILE_THRESHOLD`

Version History

- **v1.0** (Dec 2024): Separate scripts, 2010+ filter, absolute threshold
- **v2.0** (Dec 2024): Integrated workflow, 2015+ filter, percentile threshold
- **v2.1** (Dec 30, 2024): Fixed filename matching, added cleanup, species-specific thresholds

---

CONTACTS & REFERENCES

Analysis by: [Your name] **Date**: December 2024 **Related scripts**:

- `load_basemap_shapes.R` - Loads coastline, study area, WEA
- `combine_sightings_data.R` - Preprocesses and deduplicates sightings
- `process_pam_data.R` - Aggregates PAM detections by station-month

Key R packages:

- `spatstat`: KDE estimation
- `terra`: Raster processing
- `sf`: Vector operations
- `ggplot2` + `patchwork`: Visualization

End of document. Good luck, future me! 🦭