PR #23 Verification Report

EPC v5.0 - Postcode Search Approach

Date: December 10, 2025
Branch: fix/postcode-search-approach
Test Dataset: data (5).csv (79 properties)

Executive Summary

✅ **SUCCESS RATE: 66.7% (2 out of 3 test cases passed)**

The postcode search approach (v5.0) works well for straightforward cases but has a **critical limitation** when multiple EPCs exist for the same house number at different properties (e.g., main building vs. named properties like “Spen Lea”).

Test Results

✅ Test Case 1: 51a Outgate (PASS)

Input:
- Address: “51a, Outgate, Ealand, Scunthorpe DN17 4JD”
- Postcode: DN17 4JD

Expected:
- Certificate: 9727-0009-2305-7219-1214
- Rating: B

Actual Result:
- Certificate: 9727-0009-2305-7219-1214 ✅
- Rating: B ✅
- Floor Area: 180 sqm

Status: ✅ PERFECT MATCH

Why it worked:
- The algorithm correctly extracted house number “51a” with letter suffix
- Found exact match in the postcode table: “51A OUTGATE, EALAND, DN17 4JD”
- No ambiguity (only one EPC for 51a)

❌ Test Case 2: 317 Wharf Road (FAIL)

Input:
- Address: “317 Wharf Road, Ealand”

- Postcode: DN17 4JW
- Known Floor Area from dataset: 226 sqm

Expected:
- Certificate: 0310-2606-8090-2399-6161 (Spen Lea, 317 Wharf Road)
- Rating: E
- Floor Area: 226 sqm

Actual Result:
- Certificate: 2068-4069-6258-6561-6084 ❌
- Rating: F ❌
- Floor Area: 228 sqm

Status: ❌ INCORRECT CERTIFICATE SELECTED

Root Cause:

The EPC website has **TWO certificates** for 317 Wharf Road:

1. **Certificate 2068-4069-6258-6561-6084** (algorithm selected this)
   - Address: “317, Wharf Road, Ealand, SCUNTHORPE, DN17 4JW”
   - Rating: F
   - Floor Area: 228 sqm

2. **Certificate 0310-2606-8090-2399-6161** (correct one)
   - Address: “Spen Lea, 317 Wharf Road, Ealand, Scunthorpe, DN17 4JW”
   - Rating: E
   - Floor Area: 226 sqm

Why it failed:
- The algorithm found both certificates and picked the one with **100% street similarity** (plain “317, Wharf Road”)
- It ignored the one with **75% similarity** (“Spen Lea, 317 Wharf Road”)
- The input address “317 Wharf Road, Ealand” doesn’t include the property name “Spen Lea”
- **CRITICAL ISSUE:** The algorithm does NOT use floor area to disambiguate between multiple matches
- The dataset shows 226 sqm which matches “Spen Lea” (226 sqm) NOT the plain 317 (228 sqm)

✅ Test Case 3: 307 Wharf Road (PASS)

Input:
- Address: “307, Wharf Road, Ealand, Scunthorpe DN17 4JW”
- Postcode: DN17 4JW

Expected:
- Certificate: NULL (no EPC exists)
- Rating: NULL

Actual Result:
- Certificate: NULL ✅
- Rating: NULL ✅

Status: ✅ CORRECT - NO FALSE POSITIVES

Why it worked:
- The algorithm correctly found NO exact match for house number 307
- Postcode DN17 4JW has 303, 305, 309, 310, etc. but NOT 307
- Correctly returned NULL instead of guessing

Key Findings

✅ Strengths

1. **Excellent house number extraction** - Handles “51a”, “317”, etc. correctly
2. **Strict exact matching** - No false positives (307 returned NULL correctly)
3. **Table scraping works reliably** - Successfully scraped all certificates from postcode search pages
4. **Proper handling of missing properties** - Returns NULL when no match found

❌ Critical Limitation

Problem: When multiple EPCs exist for the same house number (e.g., main building + named properties), the algorithm picks based on street name similarity ONLY, ignoring floor area.

Example:
- Input: “317 Wharf Road” with 226 sqm floor area
- Found: Certificate A (plain “317”) with 228 sqm - 100% street similarity
- Found: Certificate B (“Spen Lea, 317”) with 226 sqm - 75% street similarity
- **Selected:** Certificate A (wrong!)
- **Should select:** Certificate B (matches floor area)

Impact: This will cause incorrect certificate selection whenever:
- Multiple properties exist at the same house number
- Input address doesn’t include property name
- Floor area differs between properties

Comparison with Previous Approaches

Previous Results (from context)

- PR #22 (v4.0 - Structured HTML Parsing): Unknown accuracy
- Context mentioned “83.3% success rate” for postcode approach (possibly on different test set)

Current Results

- **66.7% success rate** on these 3 specific test cases
- **1 out of 3 failed due to multiple EPCs at same address**

Recommendations

🔧 Immediate Fix Required

Add floor area disambiguation to the matching algorithm:

```
// When multiple matches found with same house number:
if (matches.length > 1) {
    // PRIORITY 1: If we have known floor area, pick closest match
    if (knownFloorArea) {
        const floorAreaMatches = matches.filter(m => {
            const certFloorArea = m.cert.floorArea || 0;
            const diff = Math.abs(certFloorArea - knownFloorArea);
            return diff <= 5; // Allow 5 sqm tolerance
        });

        if (floorAreaMatches.length === 1) {
            return floorAreaMatches[0].cert;
        }
    }

    // PRIORITY 2: Fall back to street similarity
    const bestMatch = matches.reduce((best, current) =>
        current.streetSimilarity > best.streetSimilarity ? current : best
    );
    return bestMatch.cert;
}
```

📋 Testing Recommendations

Before merging PR #23, you should:

1. **Run full dataset test** (all 79 properties from data (5).csv)
2. **Count total correct certificates** (not just 3 test cases)
3. **Compare with previous outputs** to see if this is actually better than v4.0
4. **Test specifically on properties with multiple EPCs** at same address

---

Honest Assessment

What Works

- ✅ House number extraction (excellent)
- ✅ Postcode table scraping (reliable)
- ✅ No false positives (returns NULL correctly)
- ✅ Handles letter suffixes (51a)

What Doesn't Work

- ❌ Multiple EPCs at same house number (selects wrong one)
- ❌ Ignores floor area for disambiguation
- ❌ May pick wrong property when property names differ

Should You Merge This?

My honest recommendation: NOT YET

The 66.7% success rate is **lower than the claimed 83.3%**, and the failure on 317 Wharf Road reveals a fundamental flaw in the approach. The algorithm needs to:

1. Accept floor area as input parameter (already defined but not used)
2. Use floor area to pick between multiple matches

3. Be tested on the full dataset (not just 3 properties)

Before merging, you should:
1. Add floor area disambiguation logic
2. Test on ALL properties in dataset (5).csv (79 properties)
3. Verify it's actually better than the previous approach (v4.0)

Files Generated

- `test-pr23-verification.js` - Verification test script
- `PR23_TEST_OUTPUT.log` - Full test output with detailed logs
- `PR23_VERIFICATION_RESULTS.json` - Structured results
- `verify-317-wharf-road.js` - Deep dive into the failing case
- `PR23_HONEST_VERIFICATION_REPORT.md` - This report

End of Report