

EPC Accuracy Fix - Implementation Summary

Branch: `fix/epc-accuracy`

Commit: `bd94488`

Date: December 10, 2025

Pull Request: <https://github.com/CliveCaseley/soldcomp-analyser2/pull/new/fix/epc-accuracy>

Problem Statement

Two critical EPC accuracy issues were identified in `output (65).csv` :

Issue 1: Wrong EPC Rating Extraction

- **Property:** Spen Lea, 317 Wharf Road, Ealand, Scunthorpe DN17 4JW
- **Certificate:** 0310-2606-8090-2399-6161
- **Problem:** System extracted rating **F** instead of **E**
- **Root Cause:** The dt/dd scraping method was matching “**potential rating after completing step X**” labels and extracting future improvement ratings (D, F) instead of the current rating

Issue 2: Wrong Certificate Matching

- **Property:** 307 Wharf Road, Ealand, SCUNTHORPE DN17 4JW
 - **Problem:** System matched to 303 Wharf Road’s certificate (8065-7922-4589-4034-5906)
 - **Root Cause:** 307 Wharf Road has no EPC certificate in the database, but the fuzzy matching algorithm was falling back to the closest house number (303) instead of returning null
-

Investigation Findings

Certificate Analysis

Certificate 0310-2606-8090-2399-6161 (Spen Lea 317 Wharf Road):

- Current rating: **E** (score: 45)
- Potential ratings:
- After step 1: **56 D**
- After step 2: **58 D**
- After step 3: **65 D**
- After step 4: **70 C**
- Previous bug: dt/dd method matched first “rating” label containing “potential rating after completing step 1” and extracted **D**

Certificate 8065-7922-4589-4034-5906 (303 Wharf Road):

- Address: “303, Wharf Road, Ealand, SCUNTHORPE, DN17 4JW”
- Current rating: **G**
- This was incorrectly being assigned to 307 Wharf Road

307 Wharf Road:

- **Does NOT exist** in EPC database for postcode DN17 4JW

- Available properties: 303, 305, 309, 310, 312, 314, 315, 316, 317, 318
- Previous bug: System matched to 303 (closest house number)

Implementation

Fix 1: Rating Extraction (scrapeRatingFromCertificate)

File: src/utils/epcHandler.js (lines 547-573)

Changes:

```
// Method 2: Look for "Current energy rating" or "Energy rating" in dt/dd pairs
// CRITICAL FIX: Exclude "potential rating" labels to avoid extracting future/
// potential ratings
if (!rating) {
  $('dt').each((i, elem) => {
    const label = $(elem).text().trim().toLowerCase();

    // Skip potential ratings or improvement steps
    if (label.includes('potential') ||
        label.includes('after completing') ||
        label.includes('step ')) {
      return true; // continue to next
    }

    if (label.includes('current energy rating') ||
        label.includes('energy efficiency rating') ||
        label === 'energy rating') {
      const value = $(elem).next('dd').text().trim().toUpperCase();
      // Extract letter from text like "D" or "Band D" or "Rating: D"
      const match = value.match(/\b([A-G])\b/);
      if (match) {
        rating = match[1];
        log.info(`Found EPC rating (dt/dd): ${rating}`);
        return false; // break
      }
    }
  });
}
```

Impact:

- Now skips dt labels containing “potential”, “after completing”, or “step”
- Only extracts current energy rating, not future improvement ratings
- Certificate 0310-2606-8090-2399-6161 now correctly extracts **E** instead of F/D

Fix 2: Address Matching (findBestAddressMatchFromScrapedData)

File: src/utils/epcHandler.js (lines 398-413)

Changes:

1. Raised threshold from 0.3 to 0.5:

```
// CRITICAL FIX: Require minimum score of 0.5 to ensure house number matches
// This prevents incorrect matches like 307 matching to 303
// Since house number match is weighted 70%, exact match gives 0.7 minimum
const SCORE_THRESHOLD = 0.5;
```

1. Return null instead of fallback:

```
if (bestScore >= SCORE_THRESHOLD && bestMatch) {
  // ... return match
} else {
  log.warning(`❌ No good match found (best score: ${bestScore.toFixed(3)},
threshold: ${SCORE_THRESHOLD})`);
  log.warning(`⚠️ This usually means the property doesn't have an EPC certificate`);
  log.warning(`⚠️ Returning null instead of fallback to prevent incorrect matches`);
  return null; // Changed from: return certificates[0]
}
```

File: src/utils/epcHandler.js (lines 181-186)

Changes in getCertificateNumber:

```
// CRITICAL FIX: No fallback to prevent incorrect matches
// If no good match found, return null (property has no EPC)
log.warning(`⚠️ findBestAddressMatchFromScrapedData returned null`);
log.warning(`⚠️ No EPC certificate found for this property`);
log.info('='.repeat(80));
return null; // Changed from: return fallback certificate
```

Impact:

- House number must match exactly (score 0.7+) to pass threshold 0.5
- 307 Wharf Road now returns **null** (no EPC found)
- Prevents incorrect matches like 307 → 303
- More accurate: properties without EPCs are correctly identified

Fix 3: Enhanced House Number Matching

File: src/utils/epcHandler.js (lines 262-301)

Added documentation:

```

/**
 * Calculate match score between two house number objects
 * CRITICAL FIX: Stricter matching to prevent incorrect matches (e.g., 307 to 303)
 * @param {Object} target - Target house number object
 * @param {Object} candidate - Candidate house number object
 * @returns {number} Score between 0 and 1
 */
function scoreHouseNumberMatch(target, candidate) {
  // ... existing logic ...

  // CRITICAL FIX: No partial credit for different house numbers
  // Previously, similar street names could give a match even with wrong house number
  // This caused 307 to match with 303 incorrectly
  return 0; // No match if house numbers don't match exactly
}

```

Test Results

Test Suite: test-epc-accuracy-fixes.js

All 5 tests passed:

✓ Test 1: Rating Extraction for Spen Lea 317 Wharf Road

- Certificate: 0310-2606-8090-2399-6161
- Extracted: **E** ✓
- Expected: **E** ✓

✓ Test 2: Address Matching for 307 Wharf Road

- Result: **null** ✓
- Expected: null (no EPC) ✓
- Previous bug: Would match to 303 Wharf Road

✓ Test 3: Correct Certificate for 303 Wharf Road

- Certificate: 8065-7922-4589-4034-5906 ✓
- Address: "303, Wharf Road, Ealand, SCUNTHORPE, DN17 4JW" ✓

✓ Test 4: Correct Certificate for 317 Wharf Road

- Certificate: 2068-4069-6258-6561-6084 ✓
- Address: "317, Wharf Road, Ealand, SCUNTHORPE, DN17 4JW" ✓
- Note: Correctly differentiates from "Spen Lea, 317 Wharf Road"

✓ Test 5: Rating Extraction for 303 Wharf Road

- Certificate: 8065-7922-4589-4034-5906
- Extracted: **G** ✓
- Expected: **G** ✓

Scoring Logic Summary

Address Matching Weights

- **House number match:** 70%
- **Street name match:** 30%

Score Examples

Target	Candidate	House Score	Street Score	Total	Pass?
307 Wharf Road	303 Wharf Road	0.0 (no match)	0.8	0.24	✗ No (< 0.5)
307 Wharf Road	307 Wharf Road	1.0 (exact)	1.0	1.00	✓ Yes
317 Wharf Road, Ealand	317, Wharf Road, Ealand	1.0 (exact)	1.0	1.00	✓ Yes
317 Wharf Road, Ealand	Spen Lea, 317 Wharf Road	0.0 (no house #)	0.75	0.23	✗ No

Files Changed

Modified

1. **src/utils/epcHandler.js**

- Added potential rating exclusion in `scrapeRatingFromCertificate`
- Raised threshold from 0.3 to 0.5 in `findBestAddressMatchFromScrapedData`
- Removed fallback to first certificate in `getCertificateNumber`
- Enhanced documentation for house number matching

Added

1. **test-epc-accuracy-debug.js**

- Debugging script to investigate certificate pages
- Tests multiple scraping methods
- Checks postcode search results

2. **test-epc-accuracy-fixes.js**

- Comprehensive 5-test suite
- Validates rating extraction fixes
- Validates address matching fixes
- All tests pass ✓

Deployment Instructions

Step 1: Review Pull Request

Visit: <https://github.com/CliveCaseley/soldcomp-analyser2/pull/new/fix/epc-accuracy>

Step 2: Verify Changes

```
cd /home/ubuntu/github_repos/soldcomp-analyser2
git checkout fix/epc-accuracy
node test-epc-accuracy-fixes.js
```

Expected output: All 5 tests pass

Step 3: Test with Real Data

```
# Test with problematic CSV
node src/main.js "/home/ubuntu/Uploads/data (5).csv"
```

Expected results:





- 317 Wharf Road gets rating **E** (not F)
- 307 Wharf Road gets **no EPC** (not 303's certificate)

Step 4: Merge to Master



After testing, merge the pull request to master branch.

Impact Analysis

Positive Effects

1.  **More accurate ratings:** Current ratings extracted correctly, not potential future ratings
2.  **Fewer false matches:** Properties without EPCs return null instead of wrong certificates
3.  **Better data quality:** Users can distinguish between "no EPC" and "EPC rating X"
4.  **Stricter matching:** Prevents house number confusion (e.g., 307 vs 303)

Potential Considerations

1.  **More null results:** Properties without EPCs will now show null instead of a fallback certificate
 - This is **more accurate** but may result in more blank EPC columns
 - Recommendation: Add a column "EPC Status" with values like "Not Found", "Found", "Multiple"
 2.  **Threshold change:** Raising threshold from 0.3 to 0.5 is more strict
 - May miss some valid matches if street names are significantly different
 - However, house number must still match (0.7 score from 70% weight)
 - Should be fine for most cases
-

Conclusion

Both EPC accuracy issues have been successfully fixed:

1. **Rating Extraction:** Now correctly extracts current rating E for certificate 0310-2606-8090-2399-6161
2. **Address Matching:** Now correctly returns null for 307 Wharf Road instead of matching to 303

All tests pass. The fixes improve data accuracy and prevent incorrect EPC certificate assignments.

Ready for merge and deployment.

Support & Testing

For questions or issues with this implementation:

1. Run `node test-epc-accuracy-fixes.js` to verify the fixes
2. Check logs for detailed matching scores
3. Review test cases in `test-epc-accuracy-debug.js` for debugging

End of Summary