FINCA ValiData

Using Real-Time Algorithms to Improve Field Data Collection
Who are we?

- HH Composition and Consumption
- Living Standards (Health, Education, Basic Services)
- Business Performance, Seasonality
- Employment and Job Creation
- Women’s Empowerment
- Client Aspirations

- Demand and Use of Services
- Client Satisfaction
- Competitive Positioning, Loyalty and Brand Awareness
- Branch-level Performance
- Mapping and Efficiency Analysis
Collecting Clean Data with ValiData

1. Collect data.
   - Input Prompts
   - Form Logic
   - Field Constraints
   - Filed Calculations
   - Enumerator Hints
   - Data Normalization

2. Analyze and flag problematic data.
   Observe and report on surveyor behavior.
   - Customized Validation and Outlier Check Rules (robust regressions, conditional probabilities, logistic modeling)
   - Behavioral Observations (support vector machines and random forests)

   Follow-up on flagged responses.
   Edit data.

4. Export and analyze data
   - SPSS
   - R
   - Stata
   - Excel

Call Center Field Survey

Mobile or Wi-Fi

FINCA ValiData®
Avoid Mistakes before they Happen

An Unusual Family Tree
### Pre-Fieldwork Programming

<table>
<thead>
<tr>
<th>Technique</th>
<th>PAPI, Basic CAPI</th>
<th>CAPI, ValiData</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>in-advance programming</td>
<td>X</td>
<td>Y</td>
<td>Generate new variable (e.g., consumption/household size). If per capita consumption is lower than the food poverty line, then populate the questions about food security.</td>
</tr>
<tr>
<td>skip logic</td>
<td>X</td>
<td>Y</td>
<td>If the respondent cannot read or write, skip the section on education degree.</td>
</tr>
<tr>
<td>Constraints</td>
<td>X</td>
<td>Y</td>
<td>Allow the respondent age to be between 18-89.</td>
</tr>
<tr>
<td>warning messages</td>
<td>X</td>
<td>Y</td>
<td>The respondent for this survey cannot be 2 years old. Please go back and check the response.</td>
</tr>
<tr>
<td>reminders / hints for each question</td>
<td>X</td>
<td>Y</td>
<td>Enumerator, now please switch on the GPS. Please make sure you are outside under an open sky, but very close to the interview location when capturing the GPS coordinates.</td>
</tr>
</tbody>
</table>
Avoid Mistakes…as they Happen

A Surprising Jobs Report

# Real-Time Data Quality Assessment

<table>
<thead>
<tr>
<th>ValiData Techniques</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ongoing behavioral checks using SVM and Random Forests</td>
<td>Monitor the behavior of each data collector constantly, check for possible data fabrications</td>
</tr>
<tr>
<td>high quality outlier detection</td>
<td>Robust regression models, logit/probit models,</td>
</tr>
<tr>
<td>immediate outlier correction</td>
<td>Reports the outliers and provide an immediate correction abilities as both the surveyor and the respondent are fresh and can verify</td>
</tr>
<tr>
<td>random audio audits</td>
<td>Randomly perform audio recording to check the interview quality</td>
</tr>
<tr>
<td>real time tracking — frequency, location, duration</td>
<td>Check the duration of each question/section/survey, track the path of each enumerator, monitor the progress per interviewer</td>
</tr>
</tbody>
</table>
ValiData – Detecting Falsified Surveys using SVM and RF Algorithms
ValiData – Detecting Falsified and Problematic Surveys
ValiData – Identifying Wrong Routings using SVM and RF Algorithms.
High Quality Outlier Checks

Simple Outlier Check

- Breakdown point of 0%
- Model (over) fits the data, accommodating one incorrect observation.
- Results in Type I (red) and Type II (blue) errors.

Robust Regression

- Inverse weighting on leverage
- Y-error weighted m-estimation
- Correctly identifies the outlier.
## “Outliers” vs. Outliers

<table>
<thead>
<tr>
<th>Method</th>
<th>Identified Outliers</th>
<th>Confirmed Errors</th>
<th>% correctly identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQR</td>
<td>1200</td>
<td>50</td>
<td>4%</td>
</tr>
<tr>
<td>St dev from mean/median</td>
<td>1050</td>
<td>50</td>
<td>5%</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>100</td>
<td>10</td>
<td>10%</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>50</td>
<td>48</td>
<td>96%</td>
</tr>
</tbody>
</table>

FINCA *Enterprise Study* in Pakistan
ValiData sends scheduled emails to survey managers, alerting them to specific data points which need to be reviewed. The email contains a link both the survey and the specific data field that has been flagged.

The notification includes links to the editing interface, showing the fields which require follow-up, and allowing for corrective action soon as the surveyor and the respondent are fresh and can recall the survey and the questions.

---

From: validata@fincaapps.com [mailto:validata@fincaapps.com]
Sent: Friday, April 29, 2016 12:03 AM
To: Anahit Tevosyan <Anahit.Tevosyan@finca.org>
Subject: Updates for fcatMEX_2013v5_MMP

The following surveys from survey form fcatMEX_2013v5_MMP need to be verified.

Please review the following surveys and any flagged data points within them.

After review, mark these surveys as Confirmed or Rejected.

Needs review: survey #2

- Field: MEX_D25 Date
  Problem: 840.0 violates the IQR(3) upper bound (345.0) by 495.0

Needs review: survey #8

- Field: MEX_D2 C_spent
  Problem: 122.0 violates the 3.0 standard deviation upper bound (9.905) by 112.0947

Needs review: survey #9

- Field: MEX_D3 Date
  Problem: 125000.0 violates the robust regression upper bound (400.001) by 124599.999

Needs review: survey #11

- Field: C9_me1_daysill
  Problem: The average conditional probability of selection "12" is below set threshold.

- Field: C6_me1_emp
  Problem: The logistic regression identifies "2" as non-probable response.
ValiData – Timely Editing of Data

ValiData enables correcting the outlying data points immediately and track the date, time and the object of correction. The clean data downloads are available at any point during and after the interview.
Geographic Monitoring

Track survey locations vs. sampling plan.

- Use GPS records to track the geo coverage of the field activities.
- Ensure alignment with the sampling plan
- Minimize the chances of data faking
- Enhance the enumerators’ awareness and care for their work.
Time Tracking and other Tricks

- ValiData also uses rich set of time metadata to model and crosscheck the survey accuracy including automatically generated time stamps after each question/section/module.

- Time stamp information is then used in modeling the behavior of each enumerator as well as a key variable identifying the norm duration of each question/section/module.
Real Time Quality Controls

3. May I take your picture?
   - Take Picture
   - Choose Image

1. May I record this interview?
   - Record Sound
   - Choose Sound
   - Play Sound

8. May I record you?
   - Record Video
   - Choose Video
   - Play Video

ODK Collect > Drawing Form

Gather Signature

ODK Collect > Widgets

Get Barcode

Mitch Sams
Fix Mistakes Quickly

An Uncommon Instance of Poverty
STOP Throwing Away Your Variance
The Cost Of Post Survey Outlier Cleaning

Imputations

Reduced variance, disguised genuine feedback, WRONG analysis results!!

Deletions