

Geocodificação de endereços: Visão Geral e Proposta de Avaliação de um Serviço Existente

Rafael Odon de Alencar
(odon@dcc.ufmg.br)

02/07/2009

Apresentação realizada na disciplina
Bancos de Dados Geográficos
PPGCC - DCC/UFMG

Índice

1.Introdução

I. O que é Geocodificação

II.Importância da Geocodificação

III.Técnicas de Geocodificação

IV.Fatores de Incerteza na Geocodificação

V.Geocodificação de Endereços

2.Trabalhos selecionados

3.Proposta de Avaliação de um Serviço Existente

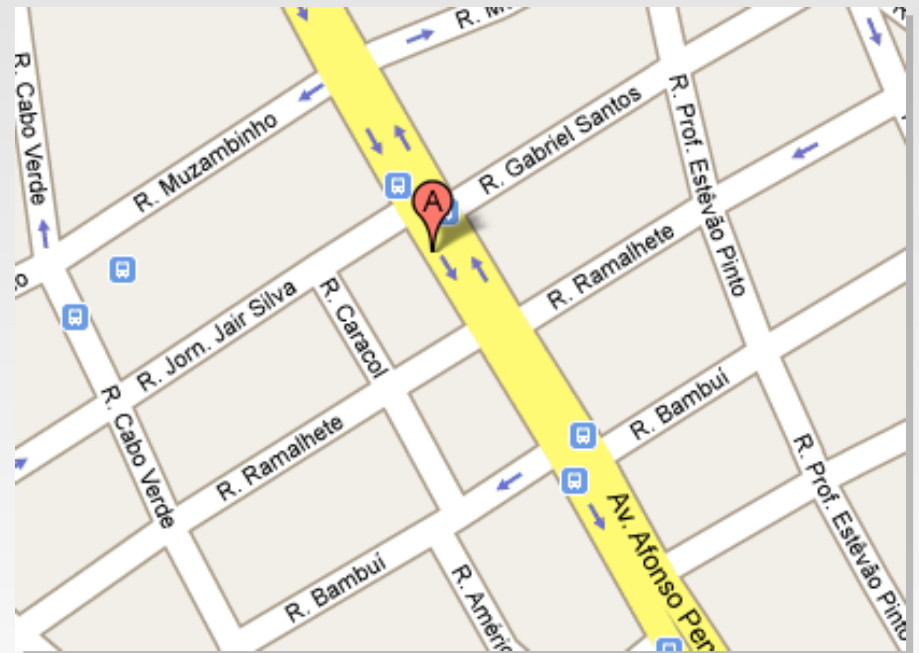
4.Conclusões

5.Referências

1. Introdução

O que é Geocodificação

- Geocodificação é o processo de associar coordenadas geográficas à locais de interesse.
- Duas visões possíveis:
 - Armazenar localizações do mundo real em um banco de dados.
 - Localizar a posição de um dado no mundo real.



Importância da Geocodificação

- É uma tarefa inicial inevitável em qualquer projeto que necessite de localização precisa de elementos no espaço do mundo real.
 - Manipulação de dados espaciais urbanos;
 - Localização de endereços de imóveis;
 - Apoio ao tráfego e logística;
 - Aplicações militares.

Técnicas de Geocodificação

- Um método direto é obter as coordenadas diretamente no mundo real, utilizando o GPS (Global Positioning System):
 - Apresenta alta precisão nos resultados;
 - Consome muito tempo;
 - Custo muito alto para grandes volumes de dados.



Técnicas de Geocodificação

- Utilização de algoritmos de geocodificação
 - Requerem um **banco de dados de referência** contendo dados já geocodificados.
 - Requer atributos que permitam **casar o dado** a ser geocodificado com os dados existentes.
- I. Primeiramente procura por uma referência no banco.
- II. Se o local existir, invoca uma técnica de **interpolação** para aproximar suas coordenadas.

Fatores de Incerteza na Geocodificação

- A **qualidade** da geocodificação pode ser afetada diretamente por dois fatores:
 - Característica da **base dados** de referência.
 - Característica da **técnica de interpolação**.

Fatores de Incerteza na Geocodificação

- Dentre as **características** do banco de dados de referência, pode-se enumerar:
 - Incompleto;
 - Com dados de casamento incorretos;
 - Com dados espaciais incorretos;
 - Inconsistente;
 - Desatualizado.

Fatores de Incerteza na Geocodificação

- Uma outra questão além da qualidade da geocodificação realizada, é a **taxa de sucesso**.
- O algoritmo pode nem sempre encontrar uma solução para o dado fornecido
 - Costuma haver um ***trade-off*** entre a precisão e taxa de sucesso;
 - Alguns algoritmos podem tentar dar algum resultados como **contróides** de áreas menos precisas.

Geocodificação de Endereços

- Amplamente utilizada.
- Tornou-se **muito popular** devido à **serviços da web** como o Google Maps, Map Quest e Yahoo Maps.
 - Criação de novas possibilidades ao integrar com o conteúdo textual da web.
- Necessita de uma **precisão muito maior** do que uma geocodificação de entidades num nível **continental**.
 - Geralmente utilizada num contexto local, para saber onde fica algum lugar de interesse.

Geocodificação de Endereços

Inclui diversas **tarefas**:

- Fazer ***parsing*** da expressão textual;
- **Padronizar** abreviaturas e notações;
- Associar cada ***token*** a uma **categoria** de casamento;
- **Indexar** as categorias necessárias;
- Procurar **candidatos**;
- Associar uma **nota** a cada candidato;
- Devolver a **melhor resposta**;

Geocodificação de Endereços

Padronização:

- A expressão fornecida pelo usuário é dividida em partes.
- As partes são avaliadas quanto ao formato (numérico, textual, abreviatura) e comparadas com dicionários para saber a qual classes podem pertencer.
 - Nome e Tipo do Logradouro, Número, Complemento, Direção (E.U.A), CEP, Cidade, Estado, etc...
- Possibilidade de utilizar técnicas probabilísticas

Geocodificação de Endereços

Casamento:

- Pode ser visto como um uma "*Linkagem*" Probabilística de Registros
 - Busca identificar a ocorrência de uma entidade em dois conjuntos distintos de dados, não diretamente associáveis.
- Aconselha-se utilizar mecanismo de blocking:
 - Poda o espaço de busca por algum critério.
 - No caso do endereços, o nome rua é o mais indicado.

Geocodificação de Endereços

Busca:

- A **indexação** é essencial para acelerar o processo.
- Pode-se utilizar a técnica ***Soundex***, que indexa baseando-se na fonética e não em sequência de caracteres.
 - Comumente aplicada à nome de rua.
- A escolha da resposta vai depender de qual **modelo de endereços** foi utilizado.

Geocodificação de Endereços

- Existem 3 modelos possíveis:
 - Rede de Ruas
 - Limites de Lotes
 - Pontos de Endereços

Geocodificação de Endereços

- Modelo de **Rede de Ruas**:
 - Referência: **segmentos** das ruas conectados por estruturas arco-nó.
 - Armazena-se o **nome** das ruas, e as **faixas de endereços** para cada segmento
 - Problema: faixa de endereços **real** ou **possível**?
 - Se o endereço desejado casar com algum segmento, realiza-se uma **interpolação linear** para localizá-lo.
 - É possível armazenar informações adicionais para retornar coordenada no **lado certo da rua** ao invés da linha central.

Geocodificação de Endereços

- Modelo de **Limites de Lotes**:
 - Referência: **delimitações dos lotes** no espaço da cidade e endereço associado a ele.
 - **Sub-endereços** estarão localizados no mesmo lugar ou não constam no banco.
 - É considerado mais preciso que a Rede de Ruas.
 - Pode gerar taxa de sucesso baixa.
 - Não é uma estrutura **própria** para geocodificação, mas pode ser aproveitada devido à disponibilidade.
 - Interesse de prefeituras em mapear os lotes.

Geocodificação de Endereços

- **Modelo de Pontos de Endereços:**
 - Cada endereço é representado por um ponto
 - Centróide do lote;
 - Centróide da construção principal do lote;
 - Local da porta principal.
 - Pode-se obter assim uma precisão de 100%, mas uma taxa de sucesso baixa para endereços ausentes na base.
 - Fuge dos falsos-positivos.

2. Trabalhos Selecionado

Trabalhos selecionados

- Foram escolhidos 2 trabalhos.
- Um deles compara 3 algoritmos de geocodificação por Redes de Ruas com coordenadas reais, afim de observar a **precisão**.
- O outro compara os 3 modelos de dados (Ruas, Lotes e Ponto) com diferentes fontes de entrada para observar a **taxa de sucesso**.

2. Trabalhos Seleccionados

Karimi e outros, 2004

*Evaluation of Uncertainties Associated with
Geocoding Techniques*

Karimi e outros, 2004

- Comparam 3 algoritmos de geocodificação de endereços:
 - **LocMatch:** proposto pelos autores, procura nos segmentos, se não achar, interpola linearmente ao longo da rua.
 - **Algoritmo do ESRI ArcView 3.2:** encontra candidatos na base, pega o mais próximo da referência interpolando linearmente ao longo da rua.
 - **Software TANA via Web:** consulta diferentes bases de referência. Se não achar o endereço, retorna centróide de área maior.

Karimi e outros, 2004

- Utilizam a base de referência TIGER:
 - Fornece dados geograficos e cartográficas de todo o **Estados Unidos**.
 - Possui estrutura de **segmentos de logradouros** indexados por CEP, com faixa de endereços possíveis.
 - Possui informações adicionais como endereços fora de sequência.

Karimi e outros, 2004

- Seleccionaram 132 endereços do TIGER.
- Coletaram **via GPS** todos os pontos, para observar a incerteza final:
 - Foram executadas 30 observações no GPS para minimizar o erro;
 - GPS acusou no máximo 3 m de erro;
 - Desvio padrão foi inferior a 4 m.
 - Coletaram pontos nos lados correto da rua, e não no meio dela;
 - Pode ainda haver influência da escolha entre ponto de porta ou outros ponto.

Karimi e outros, 2004

- Resultados:
 - Em geral os algoritmos apresentaram mais de 90% de sucesso (conseguiram achar o endereço).
 - Erro médio de aproximadamente 60 m em relação ao ponto original para todos os algoritmos.
 - Tentativas com menos de 45 m de erro:
 - 55% no TANA
 - 43% no LocMah
 - 40% no ArcView
- Conclusão: as diferentes abordagens nas técnicas de interpolação não inteferiram tanto.

2. Trabalhos Seleccionados

Zandbergen, 2007

A comparison of address point, parcel and street geocoding techniques

Zandbergen, 2007

- Selecionou algumas centenas endereços, provenientes de fontes atuais:
 - Bancos;
 - Creches;
 - Propriedades com elevadores;
 - Armazéns, supermercados e mercearias;
 - Pescadores de água salgada;
 - Criminosos sexuais.
- Todos os dados foram obtidos de órgãos, instituições, comitês e associações de interesse de cada área.

Zandbergen, 2007

- Considerou 3 regiões distintas da Florida.
- Utilizou o ArcGIS 9 para criar 3 geocodificadores baseados em diferentes bancos de dados de referência:
 - Redes de Ruas, Limites de Lotes e Pontos de Endereços.
- Configurou para ter baixa taxa de sucesso:
 - Não geocodificar buscas sem casamento direto.
- Comparou o resultado dos 3 modelos.

Zandbergen, 2007

- Taxa de sucesso da Rede de Ruas foi ligeiramente maior que a do Pontos de Endereços, que foram bem superiores ao Limites de Lote.
- Maiores taxas de sucesso para endereços residenciais em relação aos comerciais.
- Diferença substancial na taxa de sucesso das 6 fontes de dados:
 - Concluí-se que a qualidade da entrada é um fator crucial.

3. Proposta de Avaliação de um Serviço Existente

Proposta de Avaliação de um Serviço Existente

- **Proposta:** avaliar o serviço de geocodificação do Google Maps considerando como padrão de comparação a base de endereços de Belo Horizonte em pontos mantida pela PRODABEL.
- **Objetivo:** medir a precisão oferecida pelo serviço, observar regiões mais e menos afetadas, diagnosticar algumas causas de problemas.

Proposta de Avaliação de um Serviço Existente

- Utilização do web-service de geocodificação do google.



```
<code>200</code>
<request>geocode</request>
</Status>
- <Placemark id="p1">
  - <address>
    Av. Afonso Pena, 4000 - Cruzeiro, Belo Horizonte - MG, 30130-000, Brasil
  - <AddressDetails Accuracy="8">
    - <Country>
      <CountryNameCode>BR</CountryNameCode>
      <CountryName>Brasil</CountryName>
    - <AdministrativeArea>
      <AdministrativeAreaName>MG</AdministrativeAreaName>
    - <Locality>
      <LocalityName>Belo Horizonte</LocalityName>
    - <Thoroughfare>
      <ThoroughfareName>Av. Afonso Pena, 4000</ThoroughfareName>
      </Thoroughfare>
    - <PostalCode>
      <PostalCodeNumber>30130-000</PostalCodeNumber>
      </PostalCode>
    </Locality>
    </AdministrativeArea>
    </Country>
  </AddressDetails>
  - <ExtendedData>
    <LatLonBox north="-19.9434500" south="-19.9497452" east="-43.9184997" west="-43.9247949"/>
  </ExtendedData>
  - <Point>
    <coordinates>-43.9216473,-19.9465976,0</coordinates>
  </Point>
</Placemark>
</Response>
</kml>
```

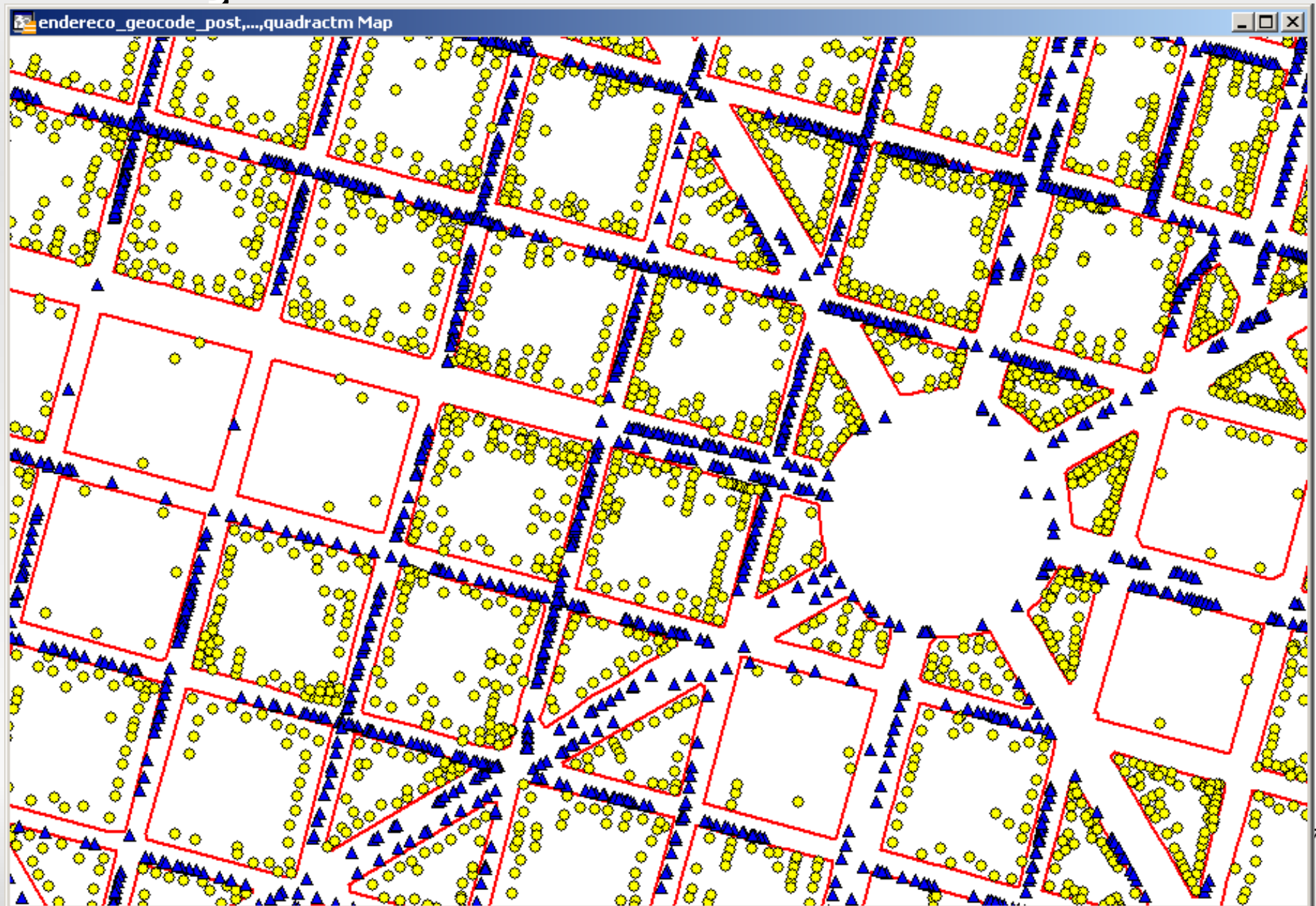
Constantes de Precisão do Serviço de Geocodificação do Google

Constante	Descrição
0	Localização desconhecida. (Desde a versão 2.59)
1	Precisão no nível do país. (Desde a versão 2.59)
2	Precisão no nível da região (estado, província, distrito etc). (Desde a versão 2.59)
3	Precisão no nível da sub-região (condado, município etc). (Desde a versão 2.59)
4	Precisão no nível da comarca (cidade, aldeia). (Desde a versão 2.59)
5	Precisão no nível do código postal (CEP). (Desde a versão 2.59)
6	Precisão no nível da rua. (Desde a versão 2.59)
7	Precisão no nível do cruzamento. (Desde a versão 2.59)
8	Precisão no nível do endereço. (Desde a versão 2.59)
9	Premissa (nome do edifício, nome da propriedade, nome do shopping etc). nível de precisão. (Desde a versão 2.105)

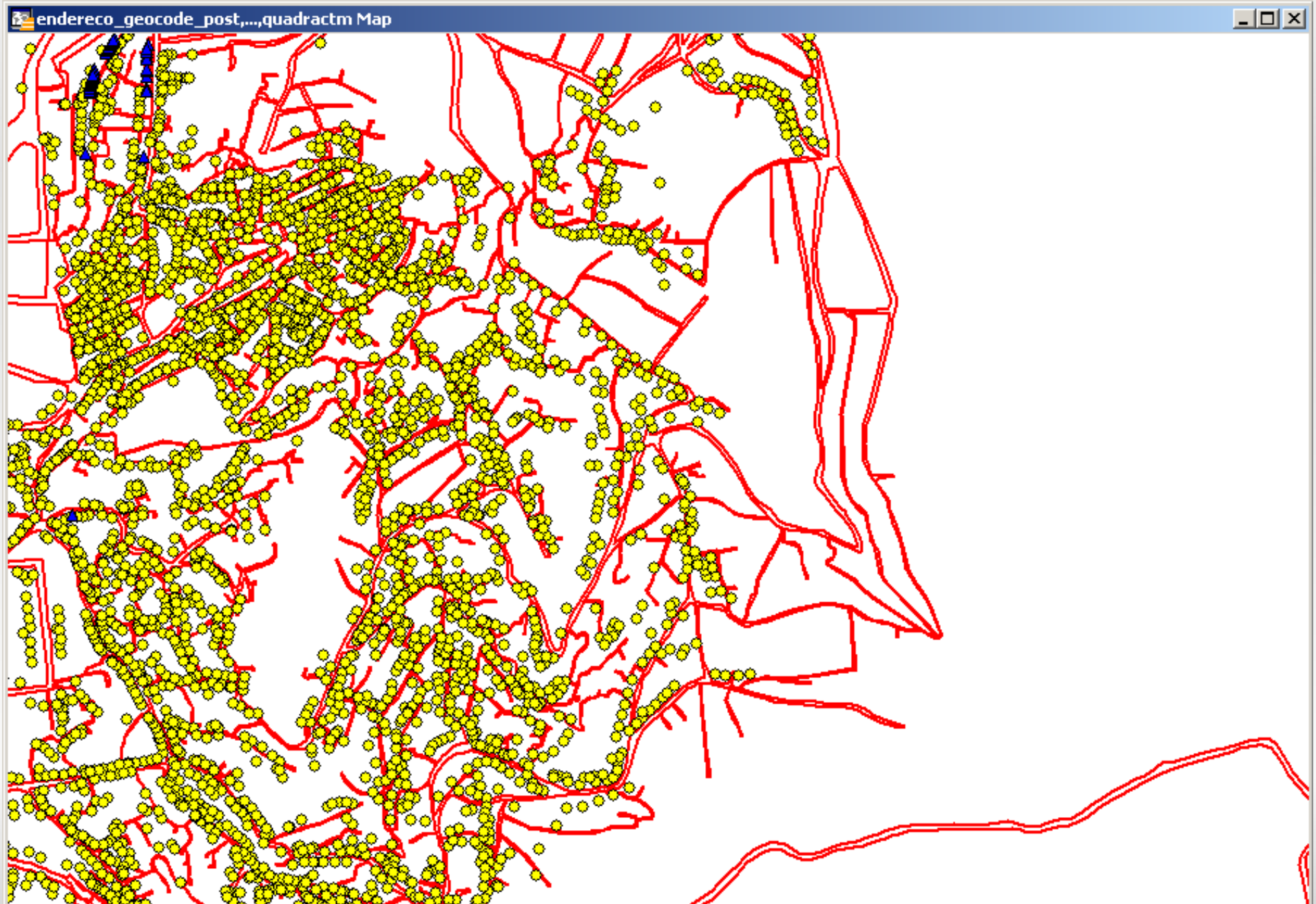
Proposta de Avaliação de um Serviço Existente

- Resultados já alcançados:
 - Obtenção da base de endereços em pontos.
 - Criação de robô em Java para comunicação com o serviço de geocodificação da Google.
 - Submissão de todos os endereços Belo Horizonte ao serviço da Google.
 - Conversão de todos os dados obtidos para o sistema de projeção UTM 23S.
 - Avaliação visual dos resultados via MapInfo.
 - Importação dos dados para o PostGIS.

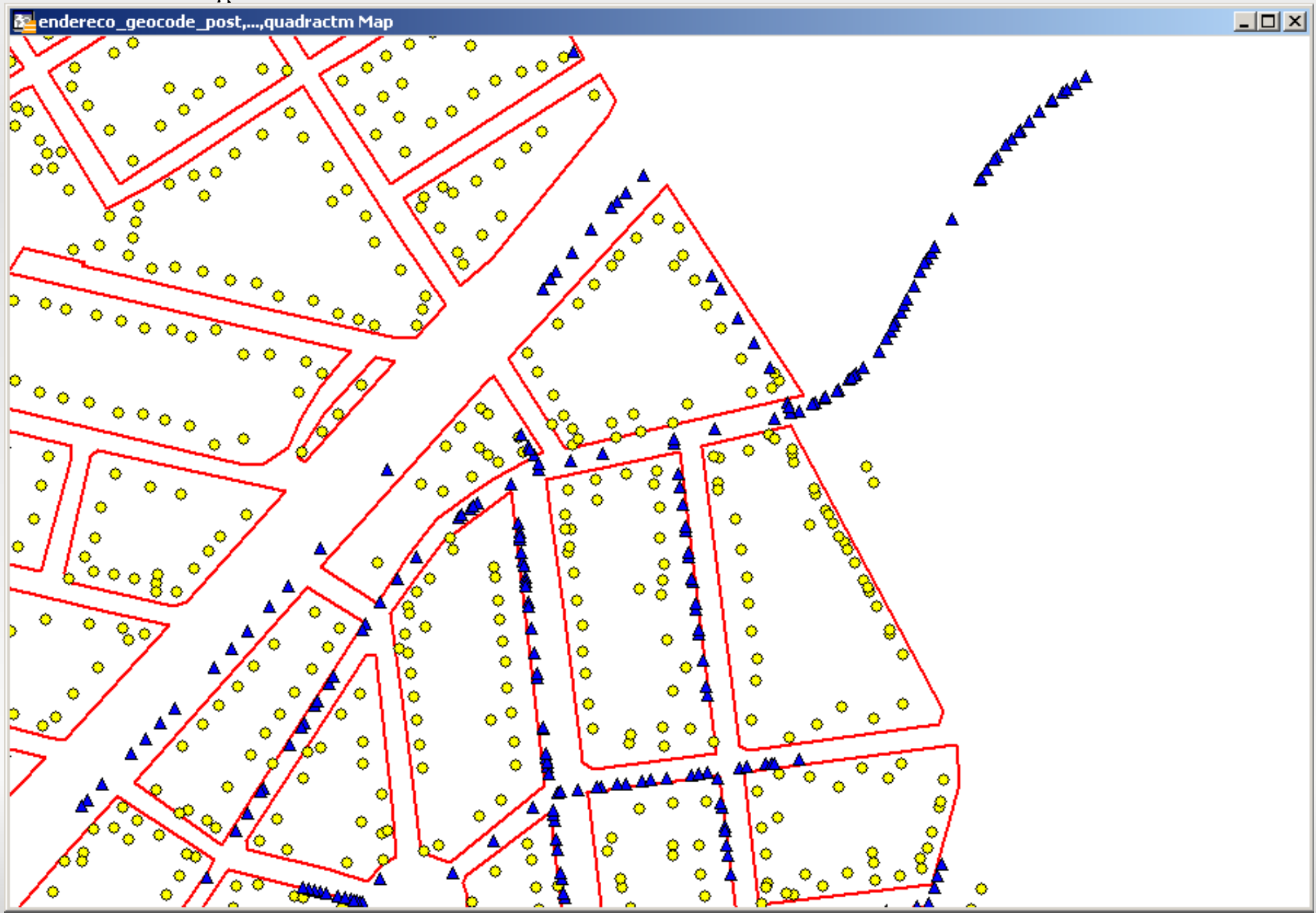
Proposta de Avaliação de um Serviço Existente



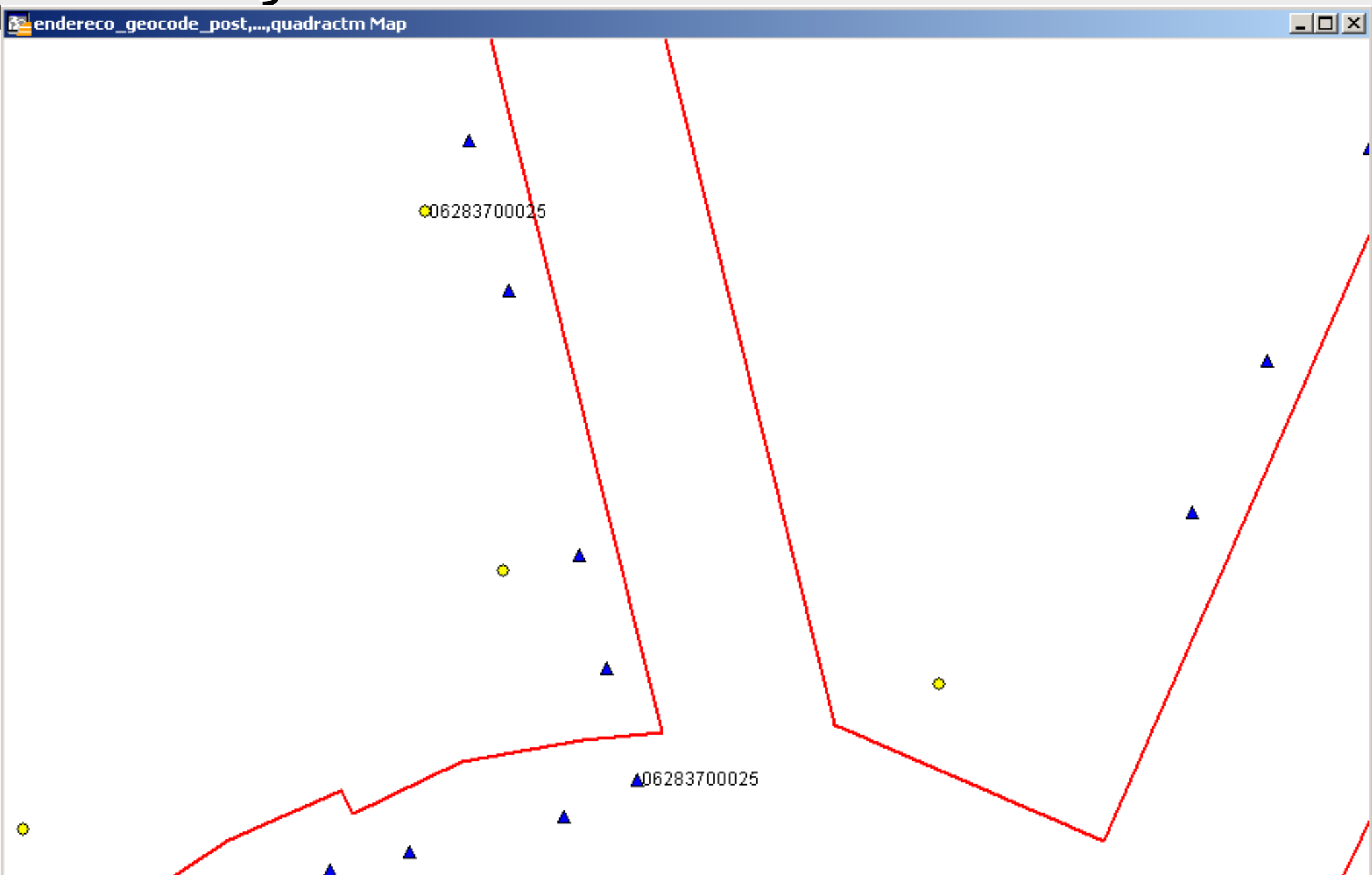
Proposta de Avaliação de um Serviço Existente



Proposta de Avaliação de um Serviço Existente



Proposta de Avaliação de um Serviço Existente



Proposta de Avaliação de um Serviço Existente

- Quantidade de endereços total: 473100
- Endereços não geocodificados (const. 0): 208.372
 - Falha ou negação de serviço ou insucesso do Google.
- Precisão no nível da rua (const. 6): 21.058
- Precisão no nível do endereço (const. 8): 243.408
- Premissa (cons. 9): 262

Proposta de Avaliação de um Serviço Existente

- Alguns dos próximos passos:
 - Avaliar a taxa de sucesso e selecionar os logradouros com alta taxa para observar a precisão neles.
 - Considerar várias medidas básicas como a distância e ângulo entre o ponto real e o ponto geocodificado.
 - Avaliar resultados em diferentes níveis: município, bairro e logradouro.
 - Diagnosticar problemas de nomenclatura (áreas sem casamento).

4. Conclusões

Conclusões

- Técnicas de geocodificação já estão estabelecidas, o grande causador de erro é a qualidade do dado de entrada e a precisão fornecida pela base de referência.
- A popularização da geocodificação de endereços na web faz com que serviços em escala global sejam amplamente utilizados no dia a dia, exigindo assim uma melhora na qualidade oferecida.
- Apesar de Belo Horizonte possuir dados de Pontos de Endereços, o Google Maps não utiliza ainda essa solução.

Referências

- **[Zandbergen, 2007]**
ZANDERBERGEN, P. A.; *A comparison of address point, parcel and street geocoding techniques.* Computers, Environment and Urban Systems. Volume 32, Issue 3, May 2008, Pages 214-232.
- **[Karimi e outros, 2004]**
KARIMI, H. A.; DURCIK M.; RASDORF W.; *Evaluation of Uncertainties Associated with Geocoding Techniques.* Computer-Aided Civil and Infrastructure Engineering. Volume 19, 2004, pages 170-185.